User Based Usability Testing

A Master Thesis by

Anders Bruun, Peter Gull and Lene Hofmeister

Aalborg University

June 2007

AALBORG UNIVERSITET

Department of Computer Science
Aalborg University
Fredrik bajers Vej 7, building E
DK-9220 Aalborg EAST
Telephone: +45 9635 8080
Fax: +45 9815 9889

Title:

Discount User Based Usability Testing

Semester:

INF8, From 1st. February 2007 to 6th. of June 2007.

Project Group:

d607a

Authors:

Anders Bruun

Peter Gull

Lene Hofmeister

Supervisor:

Jan Stage

Copies: 7
Pages: 66
Appendices: 5

Synopsis:

In this master thesis we have examined how resources spent conducting user based usability tests can be reduced. The motivation behind the thesis came from the fact that the most widely used discount usability evaluation methods do not involve users in the process. We have evaluated a method called Instant Data Analysis (IDA) and three remote asynchronous methods; User Reported Critical Incident (UCI), Forum and a longitudinal based Diary. Our findings show that the IDA method considerably reduces the efforts required to transform data into findings when conducting user based laboratory think-aloud based evaluations, while still finding the most severe problems. By using the remote asynchronous methods we were able to considerably reduce the efforts required to conduct usability tests, transform data into findings and bringing test participants to a laboratory. These methods, however, lack the ability to facilitate in the identification of serious and cosmetic problems.

Preface

This master thesis is written by three students during spring 2007 at the institute of computer science at the University of Aalborg.

We want to thank all of the participants who took the time to help us conduct all of our usability tests as well as the respondents who took the time to answer our questionnaires. A special thank goes to our supervisor Jan Stage for providing constructive feedback during the process, and to Louise and Thomas for helping us during pilot testing.

This master thesis consists of this report concerning our motivation, what we have researched and our results. The appendices consist of two articles about our main research (appendix A and B), a summary of a report concerning barriers in conjunction with usability testing in Northern Jutland (appendix C). Appendix D consists of the document used to train the participants during the remote asynchronous UCI tests. This is included for inspiration for others wanting to do similar research. A summary of the thesis can be seen in appendix E.

Contents

1 Introduction

Usability evaluation takes time, especially when conducting user-based laboratory think-aloud evaluations as described by Jeffrey Rubin [12]. According to Rubin, the following six stages have to be completed when conducting such a test:

- Develop the test plan
- Select and acquire test participants
- Prepare the test materials
- Conduct the test
- Debrief the participants
- Transform data into findings and recommendations

The most resource demanding stage of these is the transformation of data into findings in which hours of video data is thoroughly walked through to identify usability problems [7, 10]. To reduce the amount of resources, various "discount" usability methods can be applied. The most widely used discount methods are based on inspection, in which a usability expert inspects a given system for problems using a set of heuristics [2, 4, 5, 6, 8, 9, 11]. When applying inspection methods no user involvement is required, thereby giving the advantage of saving resources on the acquisition of test participants. However, the lack of user involvement is also one of the main critiques of the method, as you miss out on the problems experienced by real users.

The conduction itself of a laboratory usability test is also resource demanding because a test monitor has to present in real time during the test [1, 10, 13].

The acquisition of test participants can also be a resource demanding task to complete. We have experienced this first hand, as a local Danish company developed a system for an American customer. We conducted usability tests on the system, but it was not possible for us to do so using future users because of the geographic distance. Therefore, we were forced to use Danish participants, which required that the system dialogs and manual had to be translated into Danish. Besides from taking time, the translation also had the potential of not uncovering usability problems caused by the American formulations and use of words. The cultural background of the participants also differed from that of the future users. With the increasing use of global software development where, for instance, programming tasks are outsourced to third party companies, the same problems can potentially arise for other companies as well. When developing systems on a global scale, the bringing of test participants to a usability laboratory can be a very resource demanding task because the tests have to be conducted with the presence of test participants [3].

Thus, the stages of transforming data into findings, conducting tests and acquiring test participants require a great amount of resources to complete. Therefore, it can

be a tedious task to convince anyone that the resources required to complete these stages is worth the effort. This is the main motivation behind our master thesis. We wanted to study methods for reducing the efforts connected to usability evaluations while preserving a user-based approach.

To get further insight in the problems related to usability evaluations, we researched the barriers of applying usability testing in the software development industry. This research was done by the authors of this report along with three other researchers. A summary of the report can be seen in appendix C. 74 software development companies in Northern Jutland were asked about their experiences conducting usability tests. 39 answered. Many of them (30) stated that they conducted usability tests. These were both small and large companies. However, the respondents' understanding of usability tests differed. It was thought to be a test of functionality by 18 respondents, but 31 responded that it was a test focusing on the user in some way. An example of a more diffuse answer is: "*To get a focus group to understand the purpose of the test. We have often received feedback on design when our focus was on functionality, but we have learned from it*".

High resource usage is the number one reason, why some of these companies do not conduct usability tests. This seems reasonable as this is also seen as the main problem for those conducting usability tests. For instance, one of the respondents replied: *"The resource usage is surprisingly high"*. Another problem related to the high resource requirements was that of finding suitable test participants. Furthermore a common problem was the developers' way of thinking about usability; they can have difficulties in thinking like the users, e.g.: *"You have to think more like an ordinary user. How they would use the program"*. Some have also had problems finding motivated test participants, implementing usability tests in the development projects or making their customers see the purpose of usability testing.

Thus, the barriers experienced by the respondents are related to the high resource usage in conducting tests (and transforming data into findings) and acquiring test participants. As mentioned earlier, the most widely used discount usability evaluation methods that reduces these problems are based on inspection. These methods, however, do not invoke user involvement and have been studied in much detail in earlier research. For these reasons, we wanted to focus on alternative discount methods, which at the same time involved users. Based on the above motivations we wanted to examine the following question:

"How can the efforts spent on usability evaluations be reduced while preserving a user based approach?"

This question has led to two research questions.

1.1 Research Question 1

The time spent transforming data into findings is usually the most demanding part when conducting think-aloud user based laboratory usability tests. The first research question is therefore:

"How can the efforts spent on identification of usability problems be reduced when conducting a think-aloud user based laboratory test, and how will this affect the results?"

1.2 Research Question 2

When conducting usability evaluations, at least one evaluator has to be present in real time as test monitor during the conduction of the test. As mentioned above, the transformation of data into findings is also very resource demanding. Furthermore you have to spend resources on bringing test participants to the laboratory in order to conduct the test. We therefore wanted to examine methods that reduce these efforts, which lead to our second research question:

"Can users conduct a usability evaluation without the presence of a usability expert, and how will this affect the efforts spent and the usability problems identified?"

To answer these questions we have conducted two empirical studies, which will be presented in the following section.

2 Research Papers

This chapter presents the two research papers in this thesis. The first paper presents an evaluation of a method to reduce the time spent on analysis during user-based think aloud laboratory tests. The paper can be seen in appendix A. The second paper presents a comparison of three remote asynchronous methods for usability testing that requires the users to identify and report usability problems themselves. The paper can be seen in appendix B.

2.1 Research Paper 1

Evaluation of Instant Data Analysis – an Empirical Study

Data analysis is one of the most time consuming activities in conducting a laboratory test [7]. This paper presents a user based think-aloud laboratory experiment, where the data was analysed using two different methods; Video Data Analysis (VDA) and the discount method Instant Data Analysis (IDA). The aim of IDA is to quickly identify the most critical usability problems.

The system used for the evaluation was a healthcare system, which is intended for home use by the elderly. The participants were five elderly in the age between 61 and 78. The test took place in the usability laboratory at the university. After conducting the test, the two different analysis methods were applied:

Video Data Analysis (VDA)

> Three evaluators individually conducted a traditional Video Data Analysis, as described in [12]. The evaluators each made a problem list, and the identified problems were all categorized as either "critical", "serious" or "cosmetic". The three lists were merged into one problem list. The results from using this method was used as baseline in the experiment

Instant Data Analysis (IDA)

> The analysis took place just after the test session was finished. A data logger and the test monitor articulated and discussed the most critical problems they identified during the test. All identified problems were listed and organized on a whiteboard by a facilitator. During the evaluation the evaluators relied on their memory and the notes from the logger. The identified problems were all categorized as either "critical", "serious" or "cosmetic". At the end the facilitator wrote down all the identified problems and the severity categorizations in a problem list.

After the analysis the problem lists from VDA and IDA were merged into one.

Our results show that by using the VDA method we identified 44 problems in total, and by using the IDA method 37 problems were identified. The distribution of critical problems was 13 identified by using VDA and 16 using IDA. The two methods facilitated in the identification of the same number of serious problems,

13. 18 cosmetic problems were identified using VDA and 8 by using IDA. The three evaluators from the VDA team spent about 60 person hours on the analysis. The IDA team spent 11.5 person hours. Using two VDA evaluators instead of three do not make a significant difference in number of problems found, and the time spent is four times higher than the time spent conducting IDA.

The problem descriptions provided by IDA provide less detail than the problems described through VDA, and the latter also provides a detailed log covering the video material, which is useful in the process of redesigning the tested system. We also experienced considerable differences in the problem severity categorizations done in IDA and VDA, where the IDA categorizations generally were more severe. Finally we found that IDA was better at filtering out the potential noise created by problems experienced by one participant only.

Our results were compared with earlier research results from [7]. This comparison shows that the number of problems identified by using IDA versus problems identified by using VDA is fairly equal. The difference in person hours spent between IDA and VDA differed considerably though, as [7] used 10 times as long conducting VDA compared to IDA.

The key findings from our study show that by using IDA we were able to reduce the time spent on identification of usability problems by a factor of five, while still revealing the most severe problems. Thus, the IDA method lives up to the aim.

2.2 Research Paper 2

Comparison of Remote Asynchronous Methods for User Based Usability Testing - An Empirical Study

A laboratory think-aloud usability test requires time for conducting the test and identifying and describing usability problems. We have examined the effect of letting the users perform these tasks at home through remote asynchronous evaluation methods, without the presence of a usability evaluator. This also overcomes the problem of bringing the participants to the laboratory.

In this paper we compare three approaches to remote asynchronous usability testing. The three conditions were compared to a traditional laboratory test, which we used as a benchmark. The system used for evaluation was the e-mail client Mozilla Thunderbird, which none of the participants had used before.

We conducted a conventional laboratory user-based think-aloud test (Lab) with ten participants. The test took place in the usability laboratory at the university. The test was conducted applying the think-aloud protocol as described in [12]. To avoid being biased during the analysis, the evaluators were not present during the test.

The remote asynchronous tests were conducted using thirty different participants, ten for each condition. All participants sat at home conducting the test. They received instructions on how to conduct the test, how to identify and categorize usability problems and how to report the problems. The three conditions differed in the ways in which usability problems were reported:

User-reported Critical Incident method (UCI):

> The identified problems were all reported using an on-line form. The time spent conducting the test was e-mailed to us. The participants had a week to complete the test, but they were told to complete all tasks at one sitting.

Forum:

> The identified problems were posted on a forum and the participants were asked to discuss these. The participants were also encouraged to logon to the forum every day of the week to comment on new posts. The participants had a week to complete the test and to post and discuss in the forum. They were, however, asked to complete all tasks at one sitting. The identified problems and the time spent on conducting the test were also sent to us by e-mail.

Diary:

> The participants were asked to conduct nine tasks on the first day of the test, and the following four days the participants received two or three new tasks to complete every day, which resembled the tasks completed the first day. The participants were asked to write down the identified problems and the time spent on completing the tasks using a word processor. At the end of the test, the documents were e-mailed to us.

The data from all four conditions were collected before starting the analysis. Each dataset were given a unique identifier, and a random list was made for each evaluator. Each evaluator analysed all 40 datasets. Each evaluator made a problem list per condition (Lab, UCI, Forum and Diary), which were afterwards merged into one list of problems per condition. The four problem lists were compared and analysed and finally they were merged into one total problem list. The evaluators also categorized all problems.

The main results show that we identified 62 usability problems in total. By using the Lab we identified 46 (74%) of the total number of problems, using UCI we identified 13 (21%) of the problems, Forum 15 (24%) problems and using the Diary condition we identified 29 (47%) of all the problems. In total 21 critical problems, 17 serious and 24 cosmetic problems were identified. Using the Lab condition we identified 20 of the critical problems (95%) and using each of the asynchronous methods we identified about 50% of the critical problems. The results show that by using the Lab we identified significantly more problems than using any of the asynchronous methods. When using the asynchronous methods we generally found much less serious and cosmetic problems, with the exception of the Diary method, which facilitated in the identification of the same number of cosmetic problems as the Lab.

When looking at the total time spent on conducting the tests and identifying the problems we spent 55 person hours conducting the Lab test, 4.5 person hours applying the UCI method, 5.5 person hours applying the Forum method and 14.5 person hours applying the Diary method.

There were no significant differences in the number of problems found between the three asynchronous methods. The time spent on analysis was lowest using the UCI condition, which only required 1/12 of the time spent compared to the Lab condition. In that time we were able to identify 50 % of the critical problems.

The categorization done by the participants generally matched our categorizations, although many problems were uncategorized. UCI was the only condition in which all problems were categorized, as the on-line form required this for a problem to be submitted.

The structured approach of UCI resulted in problem reports that were easily translated into usability problem descriptions. The discussions in the forum were sparse and did not add much to the problems descriptions. The Diary condition facilitated in the identification of significantly more cosmetic problems than the other asynchronous methods, but was also the most time consuming of these. The longitudinal aspect facilitated in the identification of some extra problems, most of them being cosmetic.

The results show that it was possible for participants using the remote asynchronous conditions to identify half of the critical problems using much less time than the Lab condition. The problems were best described using the UCI method, which also facilitated in the categorization of all identified problems.

3 Research Methodology

This chapter describes the research methods, which were used to answer the two research questions, and how we used the advantages and reduced some of the disadvantages of these methods.

3.1 Description of the Research Methods

The description of the methods are based those described by Wynekoop & Conger [14]. They describe ten research methods, from which we have used two.

RQ #	Purpose of the study	Object on which the method is used	Research method	Research setting
1	Evaluation of a method	Think-aloud usability evaluation methods (both IDA and VDA)	Laboratory experiment	Artificial
2	Comparison of methods	Think-aloud usability evaluation method	Laboratory experiment	Artificial
		Remote Asynchronous usability evaluation methods	Field experiment	Natural

Table 1. Research methods used to answer the two research questions.

All the research methods have advantages and disadvantages and different purposes, which are described in the following.

Laboratory experiment

The laboratory experiments are characterized by a setting, which is created by the researcher, and the researchers have control over assignment and the set-up. It can be used to evaluate the use of a phenomenon of interest. The method assumes that real world interferences are not important. The method has the following advantages and disadvantages.

Advantages:

- High reliability
- Replicable
- Precise measures
- Great variable control
- Independent variable manipulation

Disadvantages:

- Artificial settings
- Unknown generalizability to real settings
- Assumes that real-world is not important

Field experiment

A field experiment is used for experiments, where a phenomenon is observed in a natural setting. When testing in a natural setting, it is possible to test the phenomenon in a complex social interaction and it is possible to manipulate and control the variables and to measure the changes. As the manipulation of the variables increase the naturalness of the experiments decrease. The method has the following advantages and disadvantages.

Advantages:

- Natural setting
- Replicable
- Control individual variables

Disadvantages:

- Hard to find sites
- Experiments may lose naturalness

3.2 Use of the Research Methods

In the following we describe how we have used the advantages and reduced the disadvantages of the research methods in our experiments.

Research Question 1: How can the efforts spent on identification of usability problems be reduced when conducting a think-aloud user based laboratory test, and how will this affect the results?

This research question in paper 1 covers the results of a study, where we have conducted a user-based laboratory experiment using the think-aloud protocol. Afterwards we have analysed and compared the results from the test using two different methods, VDA and IDA.

As the setting is created by the researchers, the experiment is highly replicable. The high control of the experiment has been used to make a setting comparable to that of [7], as we wanted to compare our results to theirs. Some of the variables did, however, differ from their experiment. The systems to be evaluated were not of the same type, the test participants' demographics differed and the tasks that had to be solved by the participants differed. We therefore have to assume that the VDA and IDA methods are not influenced by these variables. The similarities between our experiments and that of [7] lied in the physical set up and the data collection methods.

The VDA method may be more replicable than IDA as the researchers conducting VDA will have the data to be analysed stored on tapes, while the researchers conducting IDA will have to rely on notes taken during the test and their memory.

The artificial setting of the experiment did affect the test participants as the efforts put into solving the tasks were very high. One participant said that, had she been at home, she had put the system away instead of trying any harder to solve the current task. This also tells us that the generalizability to a real setting is limited.

One of the disadvantages of the laboratory experiment is that the participants can feel insecure and be influenced by the artificial situation. To reduce this effect the test monitor was aware of making the participants feel comfortable and described in detail the purpose of the test and the system. In the following interviews none of the participants expressed any discomfort during the course of the tests.

Research Question 2: Can users conduct a usability evaluation without the presence of a usability expert, and how will this affect the efforts spent and the usability problems identified?

In this experiment we have compared three remote asynchronous usability test methods; UCI, Forum and Diary. We chose to conduct a field experiment to get a natural setting, where the users were in their normal environments. A field experiment is close to the described normal use of the remote asynchronous usability test methods.

To evaluate the results from the remote tests, we conducted a laboratory test, which we used for comparison purposes. The laboratory setting is a controlled environment, which helped us make sure that everything worked as intended and all tests were conducted in the same way. The disadvantage of the artificial setting was reduced in the same way as the other laboratory experiment mentioned above, and only one test participant expressed discomfort during the test.

One of the advantages in conducting a field experiment is that the set-up is natural. The participants from the remote tests expressed that they liked to sit at home conducting the test. It was just like a normal situation for them. The home setting of the experiment also helped to overcome the problem of finding sites for the experiment. One of the disadvantages of our field experiment was, that we did not have as much control as in the laboratory experiment, which made it more realistic though. We cannot be certain that all participants followed our guidelines even if stated so. To reduce this effect we did a pilot test on our guidelines to make them balanced by being sufficiently detailed while not creating an overload of information.

Another problem related to the lack of control, was that three of the 30 participants did not experience any usability problem at all, which is curious since none of them have had any previous experience with the tested system. Unreliable self reporting is a problem typically associated with field studies (as opposed to field experiments). To reduce this we asked the participants to submit the time used for solving each task and gave them a hint, which enabled them to check whether or not they had solved the tasks correctly. By doing this we hoped to encourage the participants to solve all tasks.

We had great control of the demographics of the participants, how the participants were trained, which tasks the participants had to solve and how they reported problems. We used this control of the variables to change the way reporting were done between three groups of participants. The control did, however, also affect the naturalness of the experiment, as the task solving probably did not replicate the normal use of the system, as each participant might use different parts of the system during everyday use. The tasks, however, were chosen so as to only include the use of common system features.

4 Conclusion

In this master thesis we have examined how efforts spent conducting usability tests can be reduced when conducting user based usability tests. To conduct an in-depth study of this, we have set up two research questions. To answer these we have conducted two empirical studies, which concern comparative analysis of different discount user based usability evaluation methods. The two research questions and the answers to these are described in the following. Finally we answer our overall research question and present suggestions for future work

4.1 Research Question 1

"How can the efforts spent on identification of usability problems be reduced when conducting a think-aloud user based laboratory test, and how will this affect the results?"

We have evaluated the method Instant Data Analysis (IDA) and used a conventional Video Data Analysis (VDA) as a benchmark.

The key findings from our study show that we through IDA were able to reveal 68% of the total number of problems and that we found 81% of all problems using VDA. The aim of IDA is to assist in identifying the most severe usability problems in less time, and we found more critical problems using IDA than we did using VDA. We identified the same number of serious problems using IDA as we did using VDA. Additionally IDA did not facilitate in the identification of as many cosmetic problems as VDA. Using two evaluators in IDA and VDA we found that IDA required 11.5 person hours and VDA 39.25 person hours. IDA thus fulfills its purpose in revealing the most severe problems in less time than a conventional video data analysis. However, the problem descriptions provided by IDA provided less detail than the problems described through the use of VDA.

4.2 Research Question 2

"Can users conduct a usability evaluation without the presence of a usability expert, and how will this affect the efforts spent and the usability problems identified?"

We have compared three remote asynchronous methods to conduct usability testing and used a conventional laboratory think-aloud setting as a benchmark.

We found that the test participants were able to identify and report the experienced usability problems on their own. However, the participants using the Forum and Diary methods left some of the problems uncategorized.

The three remote asynchronous methods each facilitated in the identification of 50% of the total number of critical problems. Generally we found much less serious and cosmetic problems using the remote asynchronous methods. There were no significant differences in the number of problems identified between these.

Considering the fastest of the methods (UCI) we spent 1/12 of the time analyzing the results compared to the laboratory test.

4.3 Overall Research Question

"How can the efforts spent on usability evaluations be reduced while preserving a user based approach?"

When conducting usability tests, activities such as transforming data into findings, conduction of the tests and bringing test participants to a laboratory can be very resource demanding. The most widely used discount usability evaluation methods such as inspection do not invoke user involvement. In this master thesis we have evaluated alternative discount methods for conducting usability tests involving users. These methods are Instant Data Analysis and the three remote asynchronous methods: User reported Critical Incident, Forum and Diary. The four methods all reduced the efforts required to conduct user based evaluations. By using Instant Data Analysis we were able to considerably reduce the efforts required to transform data into findings. Considering the most severe problems this method performed on par with a conventional Video Data Analysis. The remote asynchronous methods considerably reduced the efforts required for transforming data into findings, conducting the test and bringing test participants to a laboratory. All of the remote asynchronous methods facilitated in the identification of half the critical problems in much less time compared to a conventional laboratory test.

4.4 Future Work

Concerning the methods examined in both articles, it would be interesting to develop these further. It would be relevant to study the usefulness of the less detailed problem descriptions provided by the IDA method, and if necessary study how to improve these. It would also be relevant to examine the effects of the user training applied during the remote asynchronous methods, and how the training might be improved to achieve more detailed problem descriptions and a higher number of identified problems.

5 Bibliography

1. Bias, Randolph G and Mayhew, Doborah J. (ed). Cost-Justifying Usability. *Acedemic Press*. 1994
2. Cockton, Gilbert and Woolrych, Alan. Sale Must End: Should Discount Methods be Cleared off HCI's Shelves?. *Interaction.*, 2002
3. Dray, Susan and Siegel, David. Remote Possibilities? International Usability Testing at a Distance. *Interactions*. 2004
4. Gray, Wayne D. Discount or Disservice? Discount Usability Analysis—Evaluation at a bargain price or simply damaged merchandise. *CHI.* 1995
5. Jeffries, Robin and Desurvire, Heather. Usability Testing vs. Heuristic Evaluation: Was there a contest?. *SIGCHI Bulletin.* 1992
6. Jeffries, Robin, Miller, James R., Wharton, Cathleen and Uyeda, Kathy M. User Interface evaluation in the real world: A Comparison of four Techniques. *Proceedings of the SIGCHI conference on Human factors in computing systems.* pages 119-124. 1991
7. Kjeldskov, Jesper et. al. Instant Data Analysis: Conducting Usability Evaluations in a Day. *Proceedings of the third Nordic conference on Human-computer interaction*. pages 233-240. 2004
8. Nielsen, Jakob. Finding Usability Problems through Heuristic Evaluation. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1992
9. Nielsen, Jakob. Usability Inspection Methods. *Conference companion on Human factors in computing systems*. 1994
10. Nielsen, Jakob. Guerilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. *http://www.useit.com/papers/guerrilla_hci.html*. 1994
11. Nielsen, Jakob and Molich, Rolf. Heuristic Evaluation of User Interfaces. *Proceedings of the SIGCHI conference on Human Factors in computing systems: Empowering people*. 1990
12. Rubin, Jeffrey. Handbook of Usability Testing. *John Wiley & Sons, INC*. 1994
13. Straub, Kath. Pitting Usability Testing against Heuristic Review. *UI Design Newsletter – September 2003.* http://www.humanfactors.com/downloads/sep03.asp
14. Wynekoop, J.L. and Conger, S.A.: A Review of Computer Aided Software Engineering Research Methods. *In Proceedings of the IFIP TC8 WG 8.2 Working Conference on The Information Systems Research Arena of The 90's*, Copenhagen, Denmark. 1990.

A. Research Paper 1

Evaluation of Instant Data Analysis – an Empirical Study

Anders Bruun, Peter Gull, Lene Hofmeister
Department of Computer Science, Aalborg University
Fredrik Bajers Vej 7, 9220 Aalborg East
anders@bruuns.org, peter@gull.dk, lene.hofmeister@gmail.com

ABSTRACT

When conducting conventional think-aloud based usability tests in a laboratory many resources (time and money) are required. One approach to overcome this while preserving the laboratory use is Instant Data Analysis (IDA). This method is based on a conventional think-aloud laboratory setting. The main idea behind IDA is to reduce the resources spent on the analysis process itself, which is the most time consuming process when conducting a conventional video based data analysis (VDA). In this paper we evaluate the IDA method in terms of problem identification and time usage. VDA is used as a benchmark. Our results show that by applying Instant Data Analysis we were able to identify 89% of the critical problems in one quarter of the time spent on VDA, through which we found 72% of the critical problems. Thus, IDA fulfills the aim of revealing the most severe problems in less time.

Keywords

Instant Data Analysis, data analysis, usability, empirical study

INTRODUCTION

For many years usability testing have been met by various barriers throughout the IT industry. These barriers are primarily grounded in the prejudice of usability testing being very resource demanding [3]. A questionnaire, which we sent out to 74 software firms in our local area in Denmark, has shown that high resource demands is the largest barrier when conducting usability tests.

In the process of making usability evaluations more widespread in the IT industry one of the first barriers to overcome is proving the return of investment [3, 11]. Upper management needs a good incentive to spend the extra money required to perform usability evaluations. Here, it is not enough to say that the end user benefits, the ultimate motivation lies in showing the return of investment in pure numbers. These numbers are almost impossible to extract since most software development projects are very diverse product wise. This makes it hard to compare projects where usability evaluations have been used with those projects where it has not been used.

Another main barrier to overcome is the cost to conduct usability evaluations [3, 16, 23]. Compared to proving the return of investment, the costs to perform usability evaluations are much easier to compare and prove, e.g. when applying two different evaluation methods on the same software. It is clear that a lower cost in conducting usability evaluations make the upper management easier to convince, and discount usability methods have proven to be constructive in getting managers in the IT industry to accept the conduction of usability evaluations within their companies [16].

It is widely acknowledged that a conventional laboratory think aloud test combined with following video analysis is the most effective in finding the greatest number of usability problems [2, 10, 12, 19, 20]. It is, however, also the most time consuming method as is produces a lot of video data that takes much time to analyze. Various discount methods exist that require less effort in analyzing the results [16].

In this article we aim to find out how the efforts spent on the analysis and identification of usability problems can be reduced when conducting laboratory usability tests, and how this will affect the results. We take a closer look at different types of discount usability evaluation methods and we evaluate one of them.

In the following we present work related to our study, next we describe the methods used for the empirical study, following this our finding are presented and finally we discuss and conclude on these.

RELATED WORK

Several usability testing methods exists that requires less effort than a conventional think aloud laboratory test. An overview of some of these is presented here.

As for discount usability, inspection is one of the most referenced [15, 17, 20] and several varieties exist (see [20] for a short overview). Instead of e.g. observing users using a system, the system is “inspected” by usability experts with the goal of unveiling potential usability problems. We here mention three methods to conduct inspection: Heuristic evaluation (HE), Cognitive Walkthrough (CW) and Inspection based on Metaphors of human thinking (MOT). HE is done by inspecting every program dialog, to see if they follow a set of usability heuristics [20]. CW, on the other hand, tries to simulate the users’ problem solving process, thereby identifying problems that the users might encounter during their process towards some goal [20]. MOT is based on five essential metaphors of human thinking, which provide usability experts with guidelines on how to consider the users’ thinking process [9].

HE is shown by [6] and [20], to facilitate the identification of more problems than CW, which is also the case when comparing MOT to CW [7]. MOT is shown to identify the same number of problems as HE [7]. [20] and [10] shows that while being more time consuming, think aloud tests tend to facilitate identification of more problems than HE. For identifying serious usability problems, think-aloud tests are even more effective than inspection, finding more serious problem per person hour used [10]. Although inspection might not facilitate in the identification of as many usability problems as other methods, it is cheap to conduct and to implement in a development process as it does not require advanced laboratory equipment and test participants.

Rapid Iterative Testing and Evaluation (RITE) tries to maintain the observation of users while lowering the effort. This method is based on a traditional think-aloud laboratory usability test. The primary focus is to make sure that identified usability problems are corrected within a short timeframe, and a secondary objective is to reduce the resources spent testing and implementing the fixes [14]. Using RITE, problems are identified on the fly. If the problems seem easy to fix they are fixed and a new prototype is used for the following tests. If a problem is not easily fixed, more data about the problem is collected during the following tests.

The RITE approach requires experienced usability experts as well as developer resources during the tests. The case study proved successful for the use in [14]. The last fixes being made were, however, tested on a small number of participants which may affect the validity, but on the other hand the fixes made were tested. The quick fixing may also affect the quality of the fixes already made, which the case study revealed, as they "broke" other parts of the user interface a couple of times during the process. The advantage of this method lies in the fact that you know that the identified usability problems are solved and that the fixes are tested too. Resource wise it is difficult to tell how effective this method is as it has not been measured. You do however avoid having to do extensive video analysis, but extra human resources are required during the test as at least one developer has to observe the test too, and a development team has to be on standby to implement fixes as problems are identified.

Additional focus on resources can be seen with Instant Data Analysis (IDA), which is also based on a conventional think-aloud test. The main idea is to reduce the time to perform analysis of user based usability testing while still identifying critical usability problems [12]. The test setting is similar to a user based think aloud laboratory test and the analysis is conducted immediately after the test sessions with the participation of the test monitor, the data logger and a facilitator. The identification of usability problems is based on the observations made by the test monitor and data logger during the test. Thus no video data analysis (VDA) has to be done.

The experiment conducted by [12] yielded good results, showing that IDA facilitated in the identification of nearly as many usability problems as VDA and only one less critical problem. This was done at only a fraction of the time taken to do VDA.

In the experiment conducted by [12] two evaluators participated in the IDA (the facilitator did not help with the identification of problems) and one evaluator in the VDA session. This unequal distribution seems unfair in the sense that it can be argued that two pairs of eyes might identify more usability problems than one pair do, making the results of the experiment biased. This issue can be presented in terms of the evaluator effect [8]. One should also bear in mind that the authors of [12] also developed the IDA method, a fact which might also have caused the results to be biased in favor of IDA.

In this article we evaluate the IDA method and use a VDA as a benchmark.

METHOD

We have conducted a conventional laboratory based think-aloud test [21], and analysed the results using two different methods:

- Video data analysis (VDA)
- Instant Data Analysis (IDA)

The results from the two methods were afterwards compared.

System

The system used for evaluation was a healthcare system (HCS) intended for home use by the elderly. It is a hardware device consisting of a display, speaker and four buttons for interaction, see figure 1. Using devices such as a blood pressure meter, a blood sugar meter or a scale the patients can perform their measurements at home and transfer these to the HCS via blue tooth, an infrared link or a serial cable. The device may also ask the patients relevant questions regarding their health. The HCS will automatically transfer the data to a nurse, doctor or whomever in charge of monitoring the patients' health. The system comes with a manual that was evaluated as well.

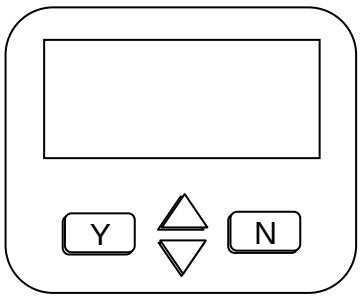

Figure 1. Sketch of the evaluated healthcare system.

Participants

The HCS was evaluated using 4 males and 1 female. Since the system primarily is intended for use by the elderly, all of our test participants were between 61 and 78 years of age. None of the participants had previous experience in using the HCS system or systems similar to it. Their experience in using electronic equipment in general varied. Two participants were novices, two were slightly practiced and the last one was experienced in using electronic equipment on a general level.

Laboratory Setting

The test was conducted in a usability laboratory; the setting is shown in Figure 2. In room 1 the test participant was sitting at a table operating the HCS. The test monitor was sitting next to the participant. Two data loggers and a technician to control cameras and microphones sat in the control room. Room 1 was equipped with cameras, a microphone and a one-way mirror.

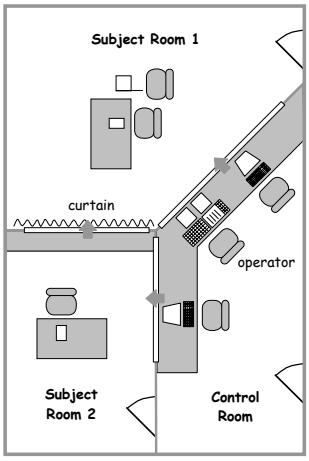

Figure 2. The setting in the usability laboratory

Figure 3. A test participant and the test monitor

Procedure

Before the test started the test participants were asked to fill out a questionnaire with demographic questions. The test monitor then introduced the participants to the system and to the evaluation procedure. This included the introduction to the think aloud procedure. The tasks were given to the test participants one by one. The test monitor's job was primarily to make sure that the test participants were thinking aloud and to give advice if the participants were completely stuck. One of the tasks was required solved, because other tasks were dependent on the result of this task. There were five tasks, and summaries of them are shown in table 1.

Task No.	Task
1	Connect and install the equipment.
2	Transfer the data from the blood sugar meter to the HCS. The blood sugar meter is connected using a cable.
3	Measure the weight and transfer the data from the scale to the HCS.
4	A new wireless blood sugar meter is used. Transfer the data from this to the HCS.
5	Clean the equipment.

Table 1. Summary of the test tasks.

Data Collection

All test sessions were recorded using video cameras and a microphone. We also used the logs written during the test sessions by two of the evaluators.

Data Analysis

The data analysis procedure was divided in two different methods, VDA and IDA. The team conducting the VDA procedure analysed the recorded video material as described below, and the IDA team conducted their analysis immediately after all the test sessions were completed, as described in [12]. The two teams did not communicate during the analysis.

VDA

Three evaluators analysed the video material individually and each made a list of identified usability problems, where every problem was categorised as either "critical", "serious" or "cosmetic".

The three lists of usability problems were discussed in the VDA-team and grouped into one list consisting of VDA identified problems only. When in doubt how to combine, split or categorize a problem, the video material was reviewed as a means to reach an agreement.

The evaluator effect (any-two agreement [8]) was calculated to 40,2%, which is above the minimum of 6% and close to the 42% maximum found in the studies of [8].

IDA

The test monitor, one of the data loggers and a facilitator conducted the IDA. We followed the approach described in [12]. The analysis was conducted using the following steps:

- The test monitor and data logger first brainstormed in 20 minutes to find any usability problems that came to mind.

- Both evaluators reviewed all the tasks, one by one, to identify more usability problems. This part lasted 30 minutes.
- The data logger reviewed the notes she took during the test for additional problems, which lasted 52 minutes.

During the evaluation the facilitator listed and organized all the identified usability problems on the whiteboard. The problems identified during the three different steps were all marked with three different colours (green, blue and black), see figure 4. By doing this it was possible later on, to identify from which step of the analysis session the usability problems were found.

After the identification the problems were categorized as either "critical", "serious" or "cosmetic". Finally the facilitator created the list of usability problems from the notes written on the whiteboard. To make sure the facilitator described the problems correctly, the list was validated and corrected by the test monitor and data logger the following day.

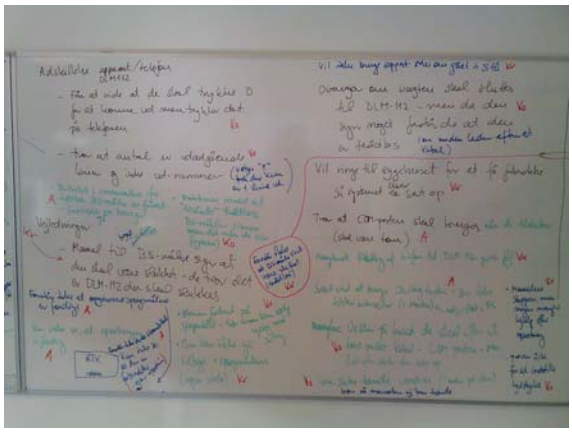

Figure 4. Picture of the whiteboard with colored notes.

Merging VDA and IDA Problem Lists

In order to compare the problem lists from the VDA and IDA procedures we had to merge these into a total list of identified usability problems. The test monitor and the data logger from IDA and the three evaluators from VDA participated in this process. In cases where the VDA and IDA lists did not have the same categorization for a particular problem, we discussed the proper categorization until everyone agreed. During the merging and the discussion, some problems were split into more problems or merged with other problems.

RESULTS

In this section we present the results from applying the VDA and IDA procedures. We start by comparing the number of problems identified and afterwards we compare the time spent on analyzing the results.

Comparison of the Number of Identified Problems

Table 2 gives an overview of the number of problems found by each method.

	VDA	IDA	Total
Critical	13	16	18
Serious	13	13	17
Cosmetic	18	8	19
Total	44	37	54

Table 2. Number of identified usability problems

In total we identified 44 usability problems using VDA. 13 of these were critical, 13 serious and 18 cosmetic. IDA revealed a total of 37 usability problems. 16 of these were critical, 13 serious and 8 cosmetic. In total we found 54 different usability problems using VDA and IDA, where 18 were critical, 17 serious and 19 cosmetic. Using the IDA method we have identified more critical problems than using the VDA method, 16 vs. 13. By using the two methods we identified the same number of serious usability problems (13). The number of cosmetic problems identified by using the VDA method (18) exceeded the number identified using IDA (8).

For both methods the majority of problems were identified during the completion of task number one, the setup of the HCS. Many of the critical problems were related to the physical setup, such as connecting cables to the correct ports but also using the HCS menu in general, e.g. issues regarding too technical terminology, missing feedback and missing information. Critical problems were also identified during connection and usage of the blood sugar meters and the Bluetooth scale. The main issues were missing feedback, too technical terminology and problems finding the correct buttons.

Applying Fishers exact test gives the value p=0.1819 for the total number of problems identified by the VDA and IDA conditions, which means that there is no significant difference. Fishers exact test gives the value p=0.418 for the critical problems identified by the two methods, no significant differences here either, and IDA identified most problems. Considering the serious problems Fishers exact test gives p=1.000 for problems identified using the VDA and IDA conditions, this means there are no differences. Comparing cosmetic problems gives us p=0.0011 using Fishers exact test, and the difference is therefore very significant. The test shows that there is no significant difference between the two methods, except when comparing cosmetic problems, which means that the IDA method meets the aim of the method; to identify the most severe problems.

Figure 5 shows the distribution of identified problems between the VDA and IDA methods. Each cell in the figure corresponds to a single problem instance. The black cells mean that the given method has identified that particular problem instance, and the white cells indicate the instances not found by that method.

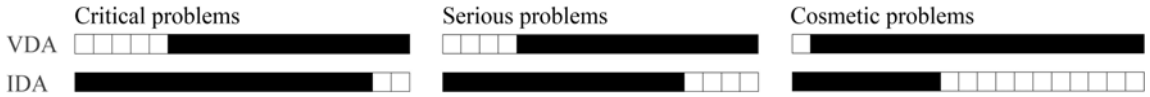

Figure 5. Problems identified using each method.

By using the IDA method we identified five critical problems, not identified using VDA, and through VDA we identified two problems, not identified applying IDA. Considering the serious problems, we identified four problems using IDA, which were not found using VDA and vice versa. We only identified one cosmetic problem, which was not found using VDA, and were able to find 11 problems using VDA, which were not identified applying IDA.

The main aim for IDA is to make an efficient identification of the most critical usability problems [12]. The results in table 2 and figure 5 show that the method lives up to this aim, since the method identified more critical problems than the VDA method, and both methods identified the same number of serious problems.

Comparison of Time Spent Analyzing

The main advantage in applying the IDA method lies in the time spent on analysis, and our results show considerable differences here.

	Eval. 1	Eval. 2	Eval. 3	Total
Identifying problems	13.5 h	13.75 h	14.5 h	41.75 h
Merging VDA lists	6 h	6 h	6 h	18 h
Total	19.5 h	19.75 h	20.5 h	59.75 h

Table 3. Time spent analyzing using the VDA method.

	Test monitor	Data logger	Facilitator	Total
IDA analysis session	2 h	2 h	2 h	6 h
Writing list of IDA problems			1.5 h	1.5 h
Validating problem list	1 h	1 h	1 h	3 h
Total	3 h	3 h	4.5 h	11.5 h

Table 4. Time spent analyzing using the IDA method.

Table 3 and table 4 give an overview of the time spent analyzing and creating the lists of problems using the VDA and IDA methods respectively. As can be seen in table 3, the total time spent conducting VDA is 59.75 person hours for the three evaluators. It should be noted that we recorded a total of four hours of video material. The time spent on IDA sums up to a total of 11.5 person hours for all three IDA participants using that method. The time spent on analysis using IDA is roughly five times lower than the time spent using VDA.

Using Two VDA Evaluators

In table 3 and table 4 VDA using three evaluators are compared to IDA using only two evaluators and a facilitator who does not identify problems.

To get a more balanced picture of VDA versus IDA we will here show how VDA, using two evaluators fairs against IDA. Table 5 gives an overview of the number of problems identified by each pair of evaluators and IDA as well as the time spent conducting the analysis.

	Eval. 1 and 2	Eval. 1 and 3	Eval. 2 and 3	All 3 eval.	IDA
Critical	13	13	12	13	16
Serious	13	12	12	13	13
Cosmetic	17	16	9	18	8
Total	43	41	33	44	37
Time spent	39.25 h	40 h	40.25 h	59.75 h	11.5 h
Problems per hour	1.1	1.0	0.8	0.7	3.2

Table 5. Problems found and time spend by all the different combinations of VDA evaluator pairs, all three VDA evaluators and the IDA evaluators.

Using only two evaluators for VDA does not do much in terms of problems found compared to IDA. An exception however is evaluator 2 and 3. They have actually found less problems overall compared to IDA.

When calculating the value of a Fishers exact test between IDA and the best-case VDA evaluator pair (evaluator 1 and 2) we get p=0.2721 for the total number of problems, and p=0.4018 for critical problems only, thus there is no significant difference.

The time spent conducting the analysis is four times as high for VDA using two evaluators compared to IDA. The number of problems identified per hour is about 3 times lower when doing IDA compared to the best case VDA evaluator pair.

The results in tables 2, 3 and 4 show that IDA is fast and effective in identifying the most critical usability problems. The results from figure 5 support this claim.

Differences in the Unique Problems Found Using One Method

Figure 5 shows how many unique problems were found using the two methods. Here we define a unique problem as a problem, which is found using one of the methods, but

not the other. In the following we examine if the identified unique problems are of a particular type of problems.

The unique critical problems found using VDA were experienced during completion of the first task, the setup of the HCS. One of the unique VDA problems is related to missing information on the HCS display. The other unique VDA problem is related to a software bug, which caused the HCS to restart during the setup process. Most of the unique critical problems identified using IDA were experienced during completion of the first task. These are problems related to the physical setup of the HCS. The fifth problem was not directly related to the system as it concerned the participants' reluctance to contact the technical support staff for help in using the HCS.

The unique serious problems found using VDA were all related to the first task. The types of problems regard the physical setup, software bugs, missing feedback and server connection errors. The unique serious problems identified using IDA were related to different tasks, and typically regarded missing feedback from the system. One of the problems was, however, not related to a particular task, but more to the overall nature of the system.

The unique cosmetic problems found using VDA are related to all tasks and varies in nature. The one unique cosmetic problem identified using IDA is related to the first task and the type of problem is regarding too technical terminology.

From the above we can see that there are no apparent differences in the types of problems, which were uniquely identified using either IDA or VDA. Also, there is no clear difference in which tasks the unique VDA and IDA problems were identified.

Problems Experienced by One Test Participant

It can be discussed whether problems experienced by a single test participant only, are generalizable or just noise [12]. Table 6 provides a base for discussing the degree of noise created by problems experienced by a single test participant. The table shows the number of unique problems identified using the VDA method only and intersecting problems identified using both the VDA and IDA method. Due to the nature of the IDA method, we have no knowledge of the number of participants experiencing each unique IDA problem.

	VDA		VDA and IDA	
	1 participant	**2 or more participants**	**1 participant**	**2 or more participants**
Critical	1	1	2	9
Serious	3	1	7	2
Cosmetic	8	3	4	3
Total	12	5	13	14

Table 6. Number of problems experienced by 1 or more participants.

As shown in table 6, 2 critical problems were identified using VDA only; 1 of these was experienced by a single participant and the other by two or more participants. From the 11 intersecting critical problems, 2 problems were experienced by a single participant, and the remaining 9 identified problems were experienced by two or more participants.

When looking at the 4 serious problems identified using the VDA method only, 3 of these were experienced by one participant, and 1 was experienced by more. Of the 9 intersecting serious problems, 7 were experienced by a single participant, and the remaining 2 problems were experienced by more.

11 of the cosmetic problems were only identified using VDA and 7 intersecting cosmetic problems were identified in the use of both methods. 8 of the 11 unique VDA problems were experienced by one test participant, and 3 problems were experienced by two or more participants. Of the 7 cosmetic problems, which are intersecting between VDA and IDA, 4 are experienced by a single participant, and the remaining 3 were all experienced by two or more participants.

The results in table 6 show that 13 out of 27 (48%) of the intersecting problems, between VDA and IDA, are experienced by a single test participant. The results also show that 12 out of 17 (70%) of the problems only identified by the use of VDA are experienced by a single participant. These results indicate that IDA is able to give the advantage of avoiding some of the noise provided by potential ungeneralizable problems. This is further discussed under the discussion section of this article.

Differences in Categorization

We experienced considerable differences in the categorizations between those problems identified during the use of VDA and IDA. Table 7 gives an overview of the initial categorizations before merging the VDA and IDA problem lists, and after the merging (in parentheses).

When merging the two lists from VDA and IDA some problems were split into multiple problems or vice versa. This explains the differences in the total number of problems before and after the merging process compared to table 2.

	VDA	**IDA**
Critical	10 (13)	17 (16)
Serious	11 (13)	12 (13)
Cosmetic	25 (18)	6 (8)
Total	46 (44)	35 (37)

Table 7. Severity categorizations before merging VDA and IDA problem lists. The numbers in the parentheses is after the merging.

Before merging the VDA and IDA lists, 49% of the identified IDA problems were categorized as critical. This

percentage was 22% for the VDA problems. When looking at the serious problems, 34% of the IDA problems and 22% of the VDA problems were categorized so. For the cosmetic problems 17% of the IDA problems were initially cosmetic compared to the 56% identified using VDA.

When looking at the original problem lists, 7 of the 35 original IDA-problems were, during the merging process, categorized to a less serious categorization and 3 to a more serious categorization. 10 of the original VDA problems were, during the merging, categorized to a more serious categorization. None were categorized to a less serious categorization.

In general the problems identified using the IDA method were categorized more seriously than the problems identified by the VDA method.

Number of Problems Identified During the Three IDA Steps

In table 8 the number of problems identified during the different steps of the IDA session are shown. 13 of the problems (37%) were identified during the brainstorm step, which lasted 20 minutes (20 % of the total time). The most time consuming step was "Reviewing the notes", in which we spent 52 minutes (51% of the time) and where 12 (34%) of the problems were identified.

	Brainstorm 20 min (20%)	**Reviewing the tasks** 30 min (29%)	**Reviewing the notes** 52 min (51%)	**Total** 102 min
Critical	7	3	7	17
Serious	4	4	4	12
Cosmetic	2	3	1	6
Total	13 (37%)	10 (29%)	12 (34%)	35

Table 8. Number of problems identified in the three IDA-steps. The numbers are before merging with the VDA problems.

The first look at the table shows no considerable difference in the number of problems identified during the "Brainstorm" and "Reviewing the notes" steps. But looking at the individual identified problems, it is revealed that some of the identified critical problems during the brainstorm were split into more detailed problems during the last step (reviewing the notes). E.g. one critical problem was split into four new critical problems when the notes were reviewed, and afterwards deleted as a "Brainstorm"-problem. This explains why the number of critical problems identified during the brainstorm is not higher. It could be expected, that this step identified the largest number, since it was the first step. The last step, even if it was the most time consuming, has shown to be important, because it contributed with important details to the already identified problems and also added new problems to the total list.

The results indicate that the combination, of brainstorm and the more structured steps, is working well. The Brainstorm seems to be a good basis for identifying the problems and the more structured steps contribute with important details and adding new problems to the problem list.

Problem Themes Identified in the Different IDA Steps

The three steps performed in the IDA session identified different problem themes. Table 9 shows the identified usability problems during the IDA analysis distributed according to the themes of problems presented in [18].

	Brainstorm	**Reviewing the tasks**	**Reviewing the notes**
Affordance		1	
Cognitive load			
Consistency			
Ergonomics			
Feedback	2	3	5
Information	6	2	2
Interaction styles		2	
Mapping		1	2
Navigation			
Task Flow			
Users' mental model	5	1	2
Visibility			1
Total	13	10	12

Table 9. Problem themes identified during the three different steps of the IDA session.

The problems identified in the "Brainstorm" step are represented in the three themes: "Users' mental model", "Information" and "Feedback".

"Information" and "Users' mental model" are the main problem themes identified during this step of the IDA session, and are mostly represented in the "Brainstorm" step compared to the other two steps. What is typical about problems of the theme "Users' mental model" is that the participants' logic is not consistent with the logic of the application. The "information" problems are mainly concerning lack of information or information that is not understandable by the user.

The problem themes in the two more structured steps; "Reviewing the tasks" and "Reviewing the notes" are more evenly distributed over all the different themes.

DISCUSSION

In this section we discuss our results and compare these to the findings of [12].

In our study we identified 89% of all critical problems using the IDA method, which is very similar to the findings in [12] where the evaluators identified 85% of all critical problems. Considering the serious problems, we

were able to find 76%, which is also similar to the 68% found in [12]. Comparing the results between VDA and IDA, [12] found that VDA facilitated in the identification of more critical problems (92%) than IDA (85%), which was opposite our study, as we identified 72% of the critical problems using VDA and 89% using IDA. The number of serious problems identified using VDA and IDA was the same in [12], which was also the case in our study.

Looking at the person hours spent, we found that the VDA method (using the fastest evaluator pair) required about 4 times more person hours than IDA. In the study conducted by [12] the VDA method required 10 times more person hours than IDA, which is a considerable difference compared to our study. In [12] there is no specific information about the steps contained in the analysis, thus the difference in person hours spent can be caused by different approaches.

Considering the unique problems experienced by one test participant, we found more of this type of problems using VDA than we did using IDA. If this type of unique problem is considered to be ungeneralizable (or noise), this can be regarded as a positive property of the IDA method. In [12] 76% of the problems only identified using VDA was experienced by one test participant only. In our study 70% of the problems only identified using VDA were experienced by one test participant only. However, 48% of intersecting problems found by both VDA and IDA in our study were also experienced by one test participant only. Thus, the IDA method is able to eliminate a high degree of the noise provided by potential ungeneralizable problems, but not all.

Before merging the VDA and IDA problem lists, the IDA problems were given a more serious severity rating than the problems found through VDA. This difference in categorization between the two methods can be caused by the fact that all IDA problems were identified based on the memory of the two evaluators. Thus, these evaluators did not have the precise information about how long time the test participants spent completing the tasks, information which the VDA evaluators had access to via the video material. Another reason can be that the test monitor experienced the user problems in a more direct manner than the video material can offer.

The level of detail in the descriptions of VDA and IDA problems differed in our study, where IDA problems were generally described in less detail than the VDA problems. E.g. a specific problem was in the IDA problem list described short, in one line. The same problem was also identified by using the VDA method and described detailed in 10 lines. The VDA problem description also contained the number of participants, who experienced the problem. This observation was also made in [12] and indicates one of the main tradeoffs when applying IDA compared to VDA. The extra person hours spent using VDA also resulted in detailed log files containing additional information on where to find examples of the problems in the video material, information which can prove vital in a redesigning process. Thus, although IDA with less effort performs on par with VDA, when looking at the number of critical and serious problems, the level of detail found applying VDA is less when using IDA.

CONCLUSIONS

In this paper we have examined how the efforts spent on identification of usability problem can be reduced when conducting laboratory usability tests and how this has affected the results.

The key findings from our study show that we by using IDA were able to considerably reduce the time spent on identification of usability problems. We were able to reveal 68% of the total number of problems using IDA and we found 81% of all problems using VDA. The aim of IDA is to assist in identifying the most severe usability problems in less time, and we found more critical problems using IDA (89%) than we did using VDA (72%). Considering the serious problems we found 76% using IDA and 76% using VDA. Using two evaluators in VDA and IDA we found that IDA required 11.5 person hours and VDA 39.25 person hours. IDA thus fulfills its purpose in revealing the most severe problems in less time than a conventional video data analysis. The problem descriptions provided by IDA, however, provide less detail than the problems described through VDA, and the latter also provides a detailed log covering the video material, which is useful in the process of redesigning the tested system. We also experienced considerable differences in the problem severity categorizations done in VDA and IDA, where the IDA categorizations generally were more severe. Finally we found that IDA was better at filtering out the potential noise created by problems experienced by one participant only. Overall we find IDA very useful in conducting discount user-based usability evaluations.

Future Work

Essentially you do not need cameras when applying the IDA method. In the future it would be interesting to study how IDA evaluation would work outside a usability laboratory.

It would also be interesting to study how useful the less detailed IDA problem descriptions are in a redesigning process.

Another interesting aspect to examine is, what the results would look like, if the steps during the IDA-evaluation were in a reverse order, e.g. if the “reviewing the tasks”-step or “reviewing the notes”–step was the first one.

REFERENCES

1. Anderson, R.E. Social impacts of computing: Codes of professional ethics. *Social Science Computing Review 10,* 2 (Winter 1992), 453-469.

2. Andreasen, Morten S. et. al. What Happened to Remote Usability Testing? An Empirical Study of Three Methods.

3. Bias, Randolph G and Mayhew, Doborah J. (ed). Cost-Justifying Usability. *Acedemic Press*. 1994.

4. CHI Conference Publications Format. Available at http://www.acm.org/sigchi/chipubform/.

5. Conger., S., and Loch, K.D. (eds.). Ethics and computer use. *Commun. ACM 38*, 12 (entire issue).

6. Desurvire, Heather W. Faster, cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing? *John Wiley & Sons, INC*. Pages 173-202. 1994.

7. Frøkjær, Erik and Hornbæk, Kasper. The Metaphors-of-Human-Thinking Technique for Usability Evaluation Compared to Heuristic Evaluation and Cognitive Walkthrough.

8. Hertzum, Morten and Jacobsen, Niels Ebbe. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human Computer Interaction 15(1)*. Pages 183-204. 2003.

9. Hornbæk, Kasper and Frøkjær, Erik. Evaluating User Interfaces with Metaphors of Human Thinking. *User Interfaces for all*. Pages 486-507. 2003.

10. Karat, Clare-Marie et. al. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. *SIGCHI conference on human factors in computing systems*. 1992.

11. Karat, Clare-Marie. Usability Engineering in Dollars and Cents. 1993.

12. Kjeldskov, Jesper et. al. Instant Data Analysis: Conducting Usability Evaluations in a Day.

13. Mackay, W.E. Ethics, lies and videotape, in *Proceedings of CHI '95* (Denver CO, May 1995), ACM Press, 138-145.

14. Medlock, Michael C. et. al. Using the RITE Method to Improve Products; a Definition and a Case Study.

15. Nielsen, Jakob. Finding Usability Problems Through Heuristic Evaluation. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1992.

16. Nielsen, Jakob. Guerilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. http://www.useit.com/papers/guerrilla_hci.html. 1994.

17. Nielsen, Jakob and Molich, Rolf. Heuristic Evaluation of User Interfaces. *Proceedings of the SIGCHI conference on Human Factors in computing systems: Empowering people*. 1990.

18. Nielsen, Cristian Monrad et. al. It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field. *NordiCHI 2006.*

19. Nielsen, Jakob. Usability Engineering. *Morgan Kaufmann.*

20. Nielsen, Jakob. Usability Inspection Methods. *Conference companion on Human factors in computing systems*. 1994.

21. Rubin, Jeffrey. Handbook of Usability Testing. *John Wiley & Sons, INC*. 1994.

22. Schwartz, M., and Task Force on Bias-Free Language. *Guidelines for Bias-Free Writing*. Indiana University Press, Bloomington IN, 1995.

23. Straub, Kath. Pitting Usability Testing against Heuristic Review. *UI Design Newsletter – September 2003.* http://www.humanfactors.com/downloads/sep03.asp.

B. Research Paper 2

Comparison of Remote Asynchronous Methods for User Based Usability Testing – an Empirical Study

Anders Bruun, Peter Gull, Lene Hofmeister
Department of Computer Science, Aalborg University
Fredrik Bajers Vej 7, 9220 Aalborg East
anders@bruuns.org, peter@gull.dk, lene.hofmeister@gmail.com

ABSTRACT
When conducting conventional think-aloud based usability tests in a laboratory many resources (time and money) are required. In this paper we have examined a branch of usability tests called remote asynchronous testing methods. The main idea behind these methods is that users complete a set of tasks at home, work or wherever appropriate. Without the presence of a usability expert the users themselves describe and report the experienced problems, hereby saving resources for conducting the user-based test and analyzing the results. In this paper we have compared three methods for remote asynchronous usability testing and used a conventional laboratory think-aloud based test (Lab) as a benchmark. The three methods are User Critical Incident Reporting (UCI), online reporting and discussion through a forum (Forum) and longitudinal user reporting through a diary (Diary). By using these three methods we have identified half the number of critical problems via each method in much less time than the Lab. By combining the results from the three asynchronous methods, we were able to identify almost the same number of critical problems as the Lab in less than half the time.

Keywords
Remote testing, asynchronous testing, usability, empirical study

INTRODUCTION
Usability testing in a traditional lab setting is known to be very demanding in terms of time and money spent on planning, conducting the test and getting the participants to the laboratory [6, 7, 18, 19, 20]. Even more resource demanding is the task of performing posttest analysis in which hours of video material showing test participants' interaction with given software or hardware systems rigorously is walked through to identify usability problems [14].

When developing for global markets the expenses for bringing test participants to a laboratory rises even further. We have experienced this first hand, as a local Danish company developed a system for an American customer. We had to conduct usability tests on the system, but it was not possible for us to do so using future users because of the geographic distance. The system dialogs and manual had to be translated into Danish, which, besides from taking time, also had the potential of not uncovering usability problems caused by the American formulations and use of words. The cultural background of the participants also differed from that of the future users.

As shown under the related work section many have done research in the field of remote usability testing, which can be divided into synchronous and asynchronous methods. In a remote synchronous setting, the test participants and evaluators are separated in space only [4]. This can be accomplished by utilizing video capture software where, for instance, the content of the test participants' screen is sent directly to the test monitor residing in a remote location [1, 2, 4, 7, 9, 11]. However, using a remote synchronous setting, the evaluators still need to be present in real time to conduct the test, and the results need to be processed by evaluators in a posttest analysis. Thus, this method is almost equally as resource demanding as a traditional lab test setting (except for participants' traveling expenses) [7].

In a remote asynchronous setting the test participants and evaluators are separated in space *and* time [4]. Applying this setting, the evaluators no longer need to be present in real time. The test participants for instance solve a number of tasks, perhaps in their own environments, using the software or hardware to be tested. As shown in the related work section, various remote asynchronous methods are applied in the literature. From a resource saving perspective, the most interesting of these remote asynchronous approaches involve those where non usability experts solve tasks and report usability problems on their own. If this is possible to accomplish, the resources spent conducting usability testing, seem to be as low as possible while still being user based.

In this article we examine if users are able to identify and report usability problems on their own in a remote asynchronous setting and what effect this has on the number of problems identified and time spent conducting tests in asynchronous conditions. This paper is structured as follows: Studies concerning asynchronous remote methods are presented under related work. The method and results from our study comparing three different remote methods are presented. In the discussion section our results are compared to those from related work after which we conclude upon our findings.

RELATED WORK

We have conducted a study of literature to reveal earlier research concerning remote asynchronous usability test. We wanted to examine earlier research, in order to determine lack in the applied methods, which could be the subject for further investigation.

The articles included here are all based on empirical studies of the use of one or more particular remote asynchronous methods for usability testing. Thus articles, which only briefly outline remote asynchronous usability testing, for instance only mentioning it under related work sections, are not included.

We have searched the following databases: ACM digital library, Article First, EBSCO, Engineering Village, Scopus, Highwire Press, IEEE Xplore, Ingenta Connect, JSTOR, Norart, SAGE Journals online, Springer Link, ISI Web of knowledge, Wiley Interscience and Google Scholar. The keywords we used in our search was: "Remote usability" and "Asynchronous usability". The relevant articles found through this search were all read, and the references made by these to other related articles were also read and included. By doing this we have identified 22 articles. Table 1 gives an overview of these.

	Remote Asynchronous Method
Traditional Lab	[1, 4, 15, 21, 22, 27, 28, 30, 31, 33, 34]
Usability Expert Inspection	[3, 4, 5, 10, 11, 25, 26, 34]
No Comparison	[8, 13, 16, 24, 32]

Table 1. Overview of identified articles in which remote asynchronous usability testing methods are applied empirically, and which comparisons are made.

As shown in table 1, 17 of these are empirical studies which compare different remote asynchronous methods. These articles each compare the results of one single remote asynchronous method with either traditional laboratory evaluation, usability expert inspection or both. [4] is the only one of the articles comparing multiple asynchronous methods. This comparison, however, is a cost-benefit comparison graph based on intuition and not empirical data. The remaining five of the 22 articles are documenting empirical studies, which apply, but do not compare the applied asynchronous methods, see [8, 13, 16, 24, 32].

In the following we describe the remote asynchronous methods applied in the 22 identified articles. We also describe how they are applied and the main results. Table 2 gives an overview of these methods.

Method:	**Article #**
Auto logging	[16, 24, 25, 26, 30, 31, 32, 33]
Interview	[8, 22, 25, 26, 31]
Questionnaires	[8, 15, 21, 24, 25, 26, 28, 30, 31, 32, 33]
User-reported Critical Incident Method	[1, 3, 4, 5, 10, 11, 27]
Unstructured Problem Reporting	[8, 15, 34]
Forum	[17]
Diary	[27]

Table 2. Methods applied for remote asynchronous usability testing.

Auto Logging, Interviews and Questionnaires

Auto logging is a method where quantitative data like visited URL history and the time used to complete tasks are collected in log files and later analysed.

The main results from [16, 24, 25, 26, 30, 31, 32, 33] indicate that this method by itself can show, e.g. if the paths to complete tasks are well designed. Although useful for that purpose the method is lacking the ability to collect the qualitative data needed to address usability issues beyond the likes of path finding and time used. This is why this method is combined with questionnaires and/or interviews as follow-ups. The results of [26] show that the auto logging method in this case found "many of the same problems" compared to the heuristic inspection. In [25] the evaluators identified 60% of the problems found via a heuristic inspection, this was, however, done over a period of about two months, which was also the case in [26]. In [33] the auto logging method is said to be "not too efficient" compared to the traditional lab setting. In both [33] and [26] there are not given any information about the total number of problems identified by the applied auto logging methods. The results of [30] show, that the evaluators identified 40% of the usability problems via the auto logging method compared to a traditional lab test setting.

User-reported Critical Incident Method

The idea behind UCI is to get users to report their experienced problems themselves. This should ideally relieve the evaluators from any work conducting tests and analysing results.

The main results from these studies show that test participants are able to report their own critical incidents. Castillo shows in [4] that a minimalist training approach works well for training participants in identifying critical incidents. There are different opinions about participants' ability to categorize the incidents. The results found by [1] indicate that participants are not good at categorizing the severity of critical incidents whereas the results from [4, 5] indicate the opposite. In this regard it should be mentioned, that the training conducted by [3, 4, 5] is more

elaborate than the one conducted by [1], which could influence the categorization results. The training conducted by [4, 5, 27] was furthermore done in physical presence with the researchers.

The total number of usability problems identified varies for the different studies. In [4] they used 24 test participants who identified 76% of the usability problems identified by experts. The results of [27] show that 10 test participants identified 60% of the problems found in a traditional lab setting. In [1] 6 non-expert test participants were able to identify 37% of the usability issues found in a traditional lab setting.

Unstructured Problem Reporting

The method of unstructured problem reporting is shortly described in [8, 15, 34]. Common for the method used in these three articles is that the participants were asked to take notes on the usability problems and other problems they encountered during completion of a set of tasks. The authors of these articles give no information about the predetermined content they wanted the participants to write down, if any. The results in [34] shows that 9 participants identified 66% of the usability problems using this kind of reporting. The researchers recommend a more structured approach to support reporting of even more usability problems, an approach more like the UCI method. The results of the study in [15] show that 8 test participants identified 50% of the total usability "issues" using the unstructured approach compared to a traditional lab setting. The results in [15] cover both negative and positive usability issues, and there is no information of how many of these reported issues are negative. The comparison done in [15] is also based on tasks, where the participants in the remote asynchronous setting solved instructional tasks and the participants during the lab setting solved exploratory tasks, an "unfair" difference which the authors themselves do take note of.

Forum

In [16] the main method used is auto logging, and the forum is used as a source for collecting qualitative data. The researchers did not encourage the participants specifically to report usability issues in the forum. Even though this was the case, the participants still reported detailed usability feedback. There is no information about user training or the number of usability problems reported in the forum. The reason can be that the purpose of [16] is not to evaluate the use of a forum for reporting usability problems per se.

[27] also addresses the issue of motivating the participants to report problems by making the reporting a collaborative effort amongst participants. The author of [27] believes that participants through collaboration may give input which increases data quality and richness compared to the UCI method. Hence, the forum seems to be a promising remote asynchronous tool for this purpose.

Diary

In [26] the primary method applied is auto logging, and diaries written by the participants on a longitudinal basis provide qualitative information. There is no information about the usefulness of the method or about the good or bad experiences with it. What is mentioned is that participants through the diaries on a longitudinal basis report on the usability problems they experience with the use of a particular hardware product. It would be interesting to experiment with the use of diaries as a standalone method to see, if the longitudinal properties will enhance participants ability to report usability problems on their own.

Concluding on the identified related articles, we have not found anyone with the purpose of comparing multiple remote asynchronous methods, except for [4], which is not based on empirical studies. Additionally, few of the papers focus on the resources required to use the methods. To answer our research question and to fill the gaps in related work, we have chosen to compare three of the seven presented asynchronous methods.

METHOD

The three asynchronous methods chosen are compared to each other using a conventional laboratory test as a benchmark. The methods chosen for comparison are:

- Laboratory testing (Lab)
- User-reported Critical Incident (UCI)
- Online reporting and discussion through a forum (Forum)
- Longitudinal user reporting through a diary (Diary)

In the rest of this section we first describe the aspects that are common for all four methods and subsequently the aspects unique to each method are described.

Participants

A total of 40 test subjects participated, ten for each condition. Half of the participants were female and the other half male. All of them studied at the University of Aalborg and were between 20 and 30 years of age. Half of them were taking a non-technical education (NT) and the other half was taking a technical education (T). For all test conditions the participants were distributed as follows: 3 NT females, 2 T females, 2 NT males and 3 T males. Most of the participants reported medium experience in using IT in general and an email client. Two participants reported themselves as being beginners to IT in general and had medium knowledge of using an email client. None of the participants had previous knowledge about usability testing.

Training

The test subjects participating in the remote asynchronous sessions were trained in the identification and categorisation of usability problems. This was done using a minimalist approach and was strictly remote and

asynchronous. All our test subjects received written instructions via email, explaining through descriptions and examples what a usability problem is, how it is identified and how it is categorised. Categorisation was divided into "low", "medium" and "high", corresponding to the traditional cosmetic, serious and critical severity categorizations [1]. Furthermore they were instructed in how to report back depending on the condition in which they participated. In general the participants have found the training material to be helpful and easy to understand.

System

We evaluated the email client Mozilla Thunderbird version 1.5, which [1] also evaluated in their work. The test participants had never used Mozilla Thunderbird.

Tasks

All participants had to solve the following tasks:

1. Create a new email account (data provided)
2. Check the number of new emails in the inbox of this account
3. Create a folder with a name (provided) and make a mail filter that automatically moves emails that has the folder name in the subject line into this folder
4. Run the mail filter just made on the emails that were in the inbox and determine the number of emails in the folder
5. Create a contact (data provided)
6. Create a contact based on an email received from a person (name provided)
7. Activate the spam filter (settings provided)
8. Find suspicious emails in the inbox, mark them as spam and check if they were automatically deleted
9. Find an email in the inbox (specified by subject line contents), mark it with a label (provided) and note what happened

We have chosen the same tasks as [1].

Laboratory Testing (Lab)

Setting

The Lab test was conducted in a usability laboratory and the setting can be seen in figure 1.

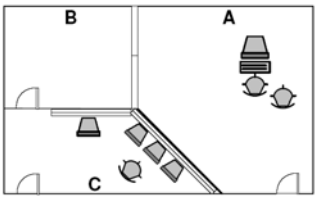

Figure 1. Overview of the usability laboratory.

In room A the test participant sat in front of the computer and next to her/him sat a test monitor whose primary task was to make sure that the test participants were thinking aloud. The room was equipped with cameras, a microphone and a one-way mirror to room C, from which the camera equipment etc. could be controlled.

Procedure

The procedure followed the guidelines of [23]. The authors did not conduct the test. The participants were introduced to the test sequence and the concept of think-aloud by the test monitor. We had scheduled one hour per participant including post-test interview and switching participants. The interviews were also done by the test monitor. The participants had to solve the nine tasks shown above, during which they had to think aloud. The test monitor was given a timeframe for the completion of each task, and had the participants not solved a task in this time, they received help from the test monitor to make sure all tasks were completed.

Data Collection

Video of the test participants' desktop was recorded along with video showing the participants' face. The test participants' and test monitor's speech were also recorded.

User-reported Critical Incident Method

Setting

In all of the remote asynchronous methods the participants sat at home using their own computer.

The participants had the possibility to do the tasks whenever they wanted; it just had to be completed before a specified date. When they started the tasks they had to finish them all at one sitting.

Procedure

The participants were asked first to go through the training material, install Mozilla Thunderbird and then begin task completion. We sent all users a guide on how to install and uninstall the program. The participants were instructed to report any negative critical incident they might find both major and minor, as soon as they discovered it. This was done using a web based report form, which we programmed using PHP, JavaScript and a MySql database. The form content was mainly the same as that used by [3] and [27]. We added the second question in the form. The following questions had to be answered using this form:

- What task were you doing when the critical incident occurred?
- What is the name of the window in which the critical incident occurred?
- Explain what you were trying to do when the critical incident occurred.
- Describe what you expected the system to do just before the critical incident occurred.
- In as much detail as possible, describe the critical incident that occurred and why you think it happened.
- Describe what you did to get out of the critical incident.
- Were you able to recover from the critical incident?

- Are you able to reproduce the critical incident and make it happen again?
- Indicate in your opinion the severity of this critical incident.

The participants were also asked to create a time log over the time spent completing each task and mail this log to us. Accompanying every task was a hint, which gave the test participants the ability to check whether or not they had solved the tasks correctly.

Data Collection

In the bottom of the online form was a submit button. When pressing this button the data was saved in an online database and the form was reset, ready to enter a new entry. The form had to run in a separate browser window, requiring the participants to shift between windows each time they encountered a problem. [10] integrated an incident reporting button directly into the tested application, which also served as a constant reminder of the reporting option. This seems like a good idea, but requires extra resources to implement. The results of [27], however, indicate that the two-windowed approach works as well.

Forum

Procedure

The participants were asked to go through the training material first and then install Mozilla Thunderbird. After installing the program the participants were asked to first take notes on the usability problems they experienced during completion of the tasks and also to categorize the severity. We encouraged them to use a word processor and not pen and paper for the note taking. They were given a list describing what we wanted them to consider when writing the problems. This list was derived from the questions used in the UCI condition. As with the UCI condition the participants were asked to finish all tasks at one sitting and to create a log over the time taken to finish each task. Accompanying every task was a hint, which gave the test participants the ability to check whether or not they had solved the tasks correctly. After completion of the tasks the test participants were instructed to uninstall Mozilla Thunderbird to keep the experiment from getting longitudinal. They were then to logon to the forum using the name and password supplied in the instructional e-mail to post and discuss their experienced usability problems with the other participants in this condition. They were given a week to post and discuss problems with each other in the forum.

When creating a new topic in the forum the participants were asked to create a subject header, which clearly expressed the core factors of the problem, thereby making it easier for other participants to see if their problems were equivalent to those already posted. Posting was allowed by all of the ten selected participants.

Each participant was given the following instructions: A) Check if the given usability problem already exists. B) If the problem does not exist then add a problem description and a severity categorization. C) If the problem is already mentioned, then comment on this either by posting an agreement with the problem description and categorization or state a disagreement with a reason indicating why either description and/or categorization is wrong.

We encouraged all participants to post and discuss the usability problems. We wanted to make posting as easy as possible, which is why we supplied every test participant with an example of how to create a post in the forum, and we also gave an example of the content, we wanted them to submit. We made sure that all participants submitted their problem descriptions and comments in an anonymous fashion. In doing so we hoped to receive even the minor usability problems, which participants, if not anonymous, would not have submitted because of embarrassment. The latter is also supported by [13].

Data Collection

The data collection procedure for this method is very simple since the forum in itself is a data collection tool. All information was written in the forum.

Diary

Procedure

The participants were, just like in the UCI and Forum conditions, asked to go through the training material first, and then install the program using the supplied installation guide. The ten participants were given a timeframe of five days to write experienced usability problems and severity categorizations in their diary. We provided the same list of elements to consider as those participating in the forum condition. They were also asked to create a time log over the time taken to finish each task. We did not force any formal content structure, as that described under UCI.

On the first day the participants received the nine tasks which were also given to all other participants. We instructed them to complete those nine tasks on the first day, and then to email the experienced problems and time log to us immediately after completion. Accompanying every task was a hint, which gave the test participants the ability to check whether or not they had solved the tasks correctly. During the remaining four days the participants received new tasks to complete on a daily basis. Because we had to compare methods, the new tasks resembled those nine tasks used in all other methods. That is, we avoided tasks that required functionality not used in the other tasks.

Data Collection

The participants were encouraged to use a word processor to write the diary and not pen and paper. As mentioned above we instructed the participants to e-mail the diary notes immediately after completion of the first nine tasks. We verified that we had received notes from all ten

participants, but did not read the notes until all data were collected from every method. After the remaining four days we received all notes on usability problems and severity categorizations from the ten participants.

Data Analysis

The data analysis was conducted by the three authors of this article. Each evaluator analyzed all data from all of the test conditions. The data consisted of 40 data sets, 10 for each of the four test conditions.

All data was collected before conducting the analysis. Each data set was then given a unique identifier, and a random list was made for each evaluator, showing the order in which to do the analysis. Each evaluator individually analyzed all the data sets one at a time.

Video Data Analysis

The video data from the laboratory test was thoroughly walked through. Every usability problem identified was described and categorized as cosmetic, serious or critical. They were also given a unique identifier to make back-tracing of the problems possible.

User-reported Data Analysis

The data from the user-reported test conditions (UCI, Forum and Diary) was read one problem at a time. Only using the information available in the users' problem descriptions, the descriptions were translated into conventional usability problem descriptions. If necessary, Mozilla Thunderbird was used to get a better understanding of the problems. A unique identifier was also added to the problem descriptions. If a user problem description could not be translated into a meaningful problem description in short time, or we could not identify the problem area using Thunderbird, the problem was not included in the problem list.

When analyzing forum descriptions, previous problem descriptions by other users in the same forum thread, was also included in the analysis, if they could contribute with anything to the description.

During the analysis we categorized the problems submitted by the users, as we wanted to make sure that categorization was done in the same way in all tests, to make a valid base for comparison. Furthermore some problems were not categorized by the users at all.

Merging of Problem Lists

When each evaluator had created a problem list for each data set, they joined the lists for each of the four test conditions (P_{1Lab}, P_{1UCI}…P_{3Diary}). These lists were then joined to form a complete problem list for each evaluator ($P_{1c,}$ $P_{2c,}$ P_{3c}). Together the three evaluators joined their individual lists for each test condition (P_{jL}, P_{jU}, P_{jF} and P_{jD}). Negotiating a joined list was done discussing each problem to reach an agreement. Categorization of problems in the joined lists was done using the most serious categorization. These joined lists for each test condition were then joined to form a complete joined problem list (P_{jc}). The joining of the lists is illustrated in figure 2. Each evaluator checked if their problems from their individual complete problem lists ($P_{1c,}$ $P_{2c,}$ P_{3c}) were present in the complete joined problem list (P_{jc}), which they were.

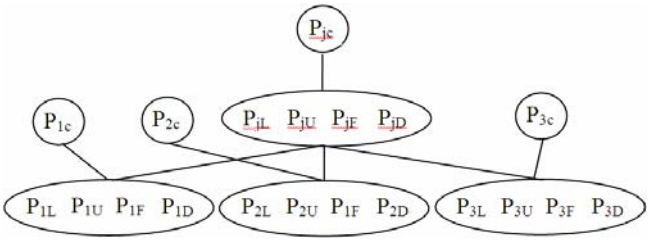

Figure 2. Joining of problem lists from each evaluator's individual problem lists for each condition (P_{1L}, P_{1U}…P_{3D}) to the complete joined problem list (P_{jc}) and complete individual problem lists (P_{ic}).

Evaluator Effect

Hertzum and Jacobsen [12] have shown that two evaluators will not find the exact same usability problems. To verify the agreement between evaluators on the usability problems, the evaluator effect has been calculated using any-two agreement. The any-two agreement shows to what extend the evaluators have identified the same problems. This has been done using equation 1 [12].

$$\text{Average of } \frac{|P_i \cap P_j|}{|P_i \cup P_j|} \text{ over all } \frac{1}{2}n(n-1) \text{ pairs of evaluators}$$

Equation 1. Calculating the evaluator effect using any-two agreement. P_i is the number of problems found by evaluator i, P_j is the number of problems found by evaluator j and n is the number of evaluators.

The evaluator effect has been calculated on the problem lists from each test method and on the combined problem lists. Table 3 shows the average any-two agreement for all of the test conditions and for the entire test.

	Lab	UCI	Forum	Diary	Avg.
Problems agreed on	23.3	9	8	17.7	14.5
Number of problems	46	13	15	29	25.8
Any-two agreement	50.7%	69.2%	53.3%	60.9%	56.3%

Table 3. The average any-two agreement between the evaluators for all test conditions.

Compared to Hertzum and Jacobsens findings [12] our any-two agreement is high. For think aloud tests their any-two agreement calculated on problem lists for three different experiments varied from 6% to 42% (avg. 18,1%, SD=20,5), whereas ours was 50,7%. Our average any-two agreement is 56,3% (SD=8,45 between the four conditions).

RESULTS

In this section we present our findings from the study. We start by presenting the main results. Next we present additional results such as task completion, task completion time, unique problems, problems found using one evaluator, differences in severity categorizations between the evaluators and test participants and finally we present the differences in the identified problems.

Comparison of Number of Problems Identified and Time Spent on Analysis

In this section we give a short overview of the main results, and subsequently we compare all the conditions with respect to the number of problems identified and the time spent performing analysis.

An overview of the problems identified can be seen in table 4. Using all four conditions we were able to identify a total of 62 usability problems. 21 of these were critical, 17 serious and 24 cosmetic.

	Lab N=10		**UCI** N=10		**Forum** N=10		**Diary** N=10	
Task completion time in minutes: Average (SD)	24.24 (6.3)		34.45 (14.33)		15.45 (5.83)		*Tasks 1-9:* 32.57 (28.34)	
Usability problems:	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Critical (21)	20	95	10	48	9	43	11	52
Serious (17)	14	82	2	12	1	6	6	35
Cosmetic (24)	12	50	1	4	5	21	12	50
Total (62)	46	74	13	21	15	24	29	47

Table 4. Number of identified problems and task completion time using the Lab, UCI, Forum and Diary methods.

Table 5 gives an overview of the person hours spent performing analysis in the Lab, UCI, Forum and Diary conditions. All timings are the sum for all three evaluators. 55 hours and 3 minutes were spent on conducting the Lab test, analysing the results and merging the problem lists from the three evaluators. The total time spent on analysis and merging of problem lists in the UCI condition was 4 hours and 33 minutes, 5 hours and 38 minutes for the Forum condition and 14 hours and 36 minutes for the Diary condition.

	Lab (46)	**UCI (13)**	**Forum (15)**	**Diary (29)**
Conducting test	10 h	0 h	0 h	0 h
Analysis	33 h 18 min	2 h 52 min	3 h 56 min	9 h 38 min
Merging problem lists	11 h 45 min	1 h 41 min	1 h 42 min	4 h 58 min
Total time spent	55 h 03 min	4 h 33 min	5 h 38 min	14 h 36 min
Avg. time per problem	1 h 12 min	21 min	23 min	30 min

Table 5. Person hours spent on conducting tests, analyzing the results and merging problem lists. The average time spent identifying each problem under the different conditions is also shown. The numbers in parentheses are the total number of problems identified under each condition.

	Lab	**UCI**	**Forum**	**Diary**
Lab		P<0.001 ***	P<0.001 ***	P=0.0031 **
UCI	P<0.001 ***		P=0.6639	P=0.002 **
Forum	P<0.001 ***	P=0.6639		P=0.0142 *
Diary	P=0.0031 **	P=0.002 **	P=0.0142 *	

Table 6. Fishers exact test for the total number of usability problems identified in the four conditions. (p)= no significant difference, * = significant difference, ** = Very significant difference, * = Extreme significant difference**

Lab

From the Lab test we identified a total of 46 usability problems. 20 of these were critical, 14 serious and 12 cosmetic. Comparing this result to the total of 62 problems, we were able to identify 74% of all problems using the Lab condition. 95% of the critical problems, 82% of the serious and 50% of all cosmetic problems were found using this method. Thus, considering the number of problems, the Lab condition was the most effective, but at the same time the most time consuming, as we spent 55 person hours conducting this test.

Lab vs. UCI

The UCI condition revealed 13 problems, 10 critical, 2 serious and 1 cosmetic. Applying Fishers exact test gives the value $p<0.001$ for the total number of problems identified between the Lab and UCI conditions, which mean that there is an extremely significant difference (see table 6 for an overview).

Using the Lab condition all the 10 critical problems also identified using UCI were identified, and they had 1 of the 2 serious problems in common. The 1 cosmetic problem identified using UCI was not found in the Lab condition. Applying Fishers exact test on each severity categorization we get $p=0.0014$ for the critical problems, $p<0.001$ for the serious problems and $p<0.001$ for the cosmetic problems. Thus, there are also significant differences when looking

amongst all the individual severity categorizations between Lab and UCI, where the Lab condition finds the highest number of problems.

The UCI condition was clearly more effective as we only spent 4½ hours identifying all problems compared to the 55 hours for the Lab condition, see table 5. Using 1/12 of the time, half the number of critical problems was found compared to the Lab condition.

Lab vs. Forum

Using the Forum condition we were able to identify a total of 15 problems. 9 were critical, 1 serious and 5 cosmetic. Fishers exact test reveals an extremely significant difference ($p<0.001$) for the total number of problems in comparing the Lab and Forum condition.

Through the Lab condition we found all 9 critical problems also identified using the Forum. The 1 serious problem from the Forum condition was also identified via the Lab and 3 of the 5 cosmetic problems were also in common between the Lab and Forum. The results from Fishers exact test show that $p<0.001$ when comparing the critical problems, $p<0.001$ for the serious and $p=0.0687$ for the cosmetic problems. There are therefore no significant difference in the identification of cosmetic problems between the Lab and Forum condition. There are extremely significant differences when looking at the critical and serious problems. Time wise we spent 5½ hours analyzing the results from the Forum condition, which is about 1/10 of the time spent on the Lab condition. Using the Lab condition the highest number of problems was identified, but the cost was much higher.

Lab vs. Diary

The Diary condition revealed a total of 29 problems, 11 critical, 6 serious and 12 cosmetic. A Fishers exact test shows a very significant difference ($p=0.0031$) in the total number of problems identified compared to the Lab.

9 of the 11 critical problems were also revealed using the Lab condition and 3 of the 6 serious problems was also in common. Finally, 3 of the 12 cosmetic problems were found using both methods. Fishers exact test show that $p=0.0036$ for the critical problems, $p=0.013$ for the serious and $p=1.000$ for the cosmetic problems. From this we can see that there is a very significant difference in the number of critical problems found by both methods and a significant difference considering the serious problems. The Lab and Diary revealed the same number of cosmetic problems.

The time spent on analysis of the Diary results was close to 14½ hours, which is about 1/4 of the time spent on the Lab condition. In that time we were able to identify little over half the number of problems found via the Lab condition. When looking at the time spent and the identification of cosmetic problems identified the Diary condition was effective.

UCI vs. Forum

The UCI and Forum conditions have 5 critical problems in common and did not find the same serious or cosmetic problems. Applying a Fishers exact test reveals no significant difference ($p=0.6639$) in the total number of problems identified between these methods.

Considering the number of problems found in the three severity categorizations, a Fishers exact test gives the value $p=1.000$ for the critical problems, $p=1.000$ for serious and $p=0.188$ for the cosmetic problems. Thus, there are no significant differences in the number of problems identified in either of the three severity categories. We spent 4½ hours on the UCI condition and 5½ hours on the Forum. Although we spent 1 hour less analyzing the UCI results, we identified 1 more critical and serious problem than we did using the Forum. On the other hand, we identified 4 more cosmetic problems through the Forum. Overall the UCI condition slightly outperforms the Forum.

UCI vs. Diary

Using the UCI and Diary conditions we found 7 critical problems and 2 serious problems common for both methods. From the total number of problems identified we found a very significant difference ($p=0.002$) when using Fishers exact test. Looking at the individual severity categorizations we calculated the values $p=1.000$ for the critical problems, $p=0.2245$ for the serious and $p<0.001$ for the cosmetic problems. From this we can see that there is no significant difference in the number of critical and serious problems identified using UCI and Diary. There is, however, an extremely significant difference when looking at the cosmetic problems. The Diary condition facilitated in identification of the highest number of problems in all three severity categories but the time spent was 14½ hours versus 4½ hours for the UCI condition. When taking the time spent into consideration we find the UCI method more efficient than the Diary.

Forum vs. Diary

Through the Forum and Diary conditions we found 7 critical problems, 1 serious and 1 cosmetic problem common for both methods. Considering the total number of problems identified we found a significant difference ($p=0.0142$) between these methods using Fishers exact test. Focusing on the three severity categorizations we get $p=0.7578$ for the critical problems, $p=0.0854$ for serious and $p=0.0687$ for the cosmetic problems using Fishers exact test. Thus, there is no significant difference between the number of critical, serious or cosmetic problems individually, which are identified via the Forum and Diary conditions. We did, however, spend 14½ hours on the Diary condition compared to 5½ hours on the Forum condition.

Task Completion

For all 40 test participants the mean value of completed tasks is 8.9, and the standard deviation is 0.2. The only

condition, in which not all tasks were completed, was UCI. Here one test participant did not complete tasks 3 and 4. All participants completed the 9 tasks in the Lab condition, but the majority of participants experienced difficulties in completing tasks 3, 6 and 7. It should be noted that the help from the test monitor caused all Lab test participants to complete all tasks. Problems for which participants received help in the completion were, however, all categorized as critical.

Task Completion Time

Table 4 gives an overview of the time spent completing all tasks. Considering tasks 1-9 the most significant difference is between the Forum and UCI conditions. Participants using the Forum spent an average of 15.45 (SD=5.83) minutes completing all 9 tasks and UCI participants spent 34.45 (SD=14.33) on average. In between these we find the Lab condition, in which participants spent an average of 24.24 (SD=6.3) minutes and the Diary with a 32.57 (SD=28.32) minute average.

Considering the standard deviations, there is a considerable difference in the participants' completion time for the Diary condition compared to the other conditions. The completion times varied from a minimum of about 4 minutes to complete tasks 1-9 up to a maximum of 99 minutes.

Unique Problems

Through the different test conditions we have identified different problems only found in one condition. Furthermore some of the participants have experienced problems not found by others. In this section we will present our findings of these unique problems.

Problems Identified in One Test Condition Only

Having revealed usability problems not revealed by other methods can tell us more about the uniqueness and strengths of a particular test condition. Table 7 gives an overview of the number of problems identified in one test condition only and in which condition, these are identified.

	Lab	UCI	Forum	Diary	Total
Critical (21)	5	0	0	1 / **1**	6
Serious (17)	11	1	0	2 / **0**	14
Cosmetic (24)	7	0	2	9 / **3**	18
Total (62)	23	1	2	12 / **4**	38

Table 7. The number of problems identified during one test condition only. The numbers in parentheses are the total number of problems for each categorization and the numbers in bold are the number of unique problems identified using the diary during the extra days of task solving.

From table 7 it is clear that the Lab test revealed many problems not found by any of the remote asynchronous conditions. 37% percent of the problems were identified only using the Lab condition. The majority of these are serious and 24% of all critical problems identified are only identified using the laboratory condition. Looking at all three severity categories, the unique Lab problems are primarily of the theme "Information" defined in [17], e.g. problems in which the participants were missing information or that the given information from the system was too technical to understand. The UCI and Forum conditions, also being the ones revealing the smallest number of problems in total, have revealed 3 unique problems in total, not one of them being critical. The diary however, has revealed even more unique cosmetic problems than the laboratory condition (9). This is, however, not because of the extra tasks that the participants had to solve, as the extra days spent by the participants solving tasks in Thunderbird revealed a total of four extra problems, three of them being cosmetic. The unique problems found via the Diary condition are distributed evenly over the different problem themes defined in [17]. The Lab and Diary are the two methods from which we identified most of the unique problems.

It is also very interesting that we through the Lab condition found 5 critical problems not found by any of the remote asynchronous methods. This means that by combining the results from the UCI, Forum and Diary conditions we almost found all critical problems. The total time spent on analyzing all three asynchronous conditions sums up to about 24 person hours (see table 5), which is about half the time spent analyzing the Lab test results.

Problems Experienced by One Test Participant Only

Problems identified by one test participant only may be considered noise as it can be argued that these cannot be generalized [14]. Table 8 gives an overview of such problems distributed over the four conditions.

	Lab	UCI	Forum	Diary	Total
Critical (21)	3	0	0	0 / **0**	3
Serious (17)	6	1	0	2 / **0**	9
Cosmetic (24)	6	0	2	4 / **1**	12
Total (62)	15	1	2	6 / **1**	24

Table 8. The number of problems experienced by one test participant only. The numbers in parentheses are the total number of problems for each categorization and the numbers in bold are the number of unique problems identified using the diary during the extra days of task solving.

As shown in table 8, 24 of the 38 unique problems (63%) from table 7 have only been revealed during one test session. That includes all of the unique problems identified using UCI and Forum, as well as half of the problems identified using diary and 65% of the problems identified during the laboratory condition (see table 7). If we consider these problems to be noise, the actual number of usability problems is 38, a 39% drop. It is primarily the serious and cosmetic problems that are affected by this. The drop in critical problems will be 14%. From table 8 it is clear, that the asynchronous conditions are much better at filtering out the unique problems identified by one participant.

Using One Evaluator

In this experiment there have been three evaluators. The amount of resources required for three evaluators to conduct usability evaluations similar to the ones we have done will in many cases be too excessive. For this reason, to illustrate "the real world", we have selected the results of the worst-case evaluator, the one who identified the lowest number of problems. This is compared to the results from all three evaluators.

Using a Fishers exact test shows a very significant difference in the total number of problems found in the Lab condition comparing the worst case evaluator to the three evaluators. Additionally we found no significant difference in the total number of problems when doing the same comparison on the asynchronous conditions.

The above evaluator case is an interesting observation in the sense that the asynchronous conditions in our case reduces the evaluator effect to a point where there is no significant difference in the number of problems identified, which is not the case with the Lab condition.

The person hours spent is also minimized, and there is no need to merge the problems lists from multiple evaluators, which we spent 1½ person hours doing for the UCI list, 2 person hours for the Forum list and 6 person hours for the Diary list.

Differences Between Participant and Evaluator Categorizations

All the users had received the same instructions on how to categorize the usability problems. In this section we examine whether it was possible for the test participants from the remote tests to categorise the problems properly. Table 9 shows the number of categorizations, which did or did not match the categorizations done by the evaluators.

	UCI (13)		**Forum (15)**		**Diary (29)**	
	#	%	#	%	#	%
Same categorisation	10	77	10	66	13	45
No categorisation	0	0	4	27	11	38
Lower participant categorisation	1	8	1	7	2	7
Higher participant categorisation	2	15	0	0	3	10
Total	13	100	15	100	29	100

Table 9. Number of problems, where the participants' categorisations did or did not match those done by the evaluators.

The most structured method of the three, UCI, was the one which categorizations matched the categorization made by the evaluators the best. In this case 3 of the 13 found problems (23%) did not match. All identified problems in the method were categorized, and this was due to the fact that it was not possible to report a problem without a categorisation.

As can be seen from table 9, the Diary condition was the one method where 11 out of 29 (38%) problems were uncategorized.

The Forum method is, like the Diary, a more unstructured approach than UCI. In this case 5 problems (34%) were not categorized correctly, which also includes problems without any categorization.

The most significant differences shown in table 9 are problems without any categorization. Here 15 problems were not categorized at all, which is an issue only occurring with the use of the unstructured Forum and Diary conditions where no formal structure is forced upon the participants. These results thus indicate that the UCI condition leaves less extra work for the evaluators to perform afterwards.

Difference in Problems

We have seen a difference in the problem themes identified using the different methods.

The critical problems identified using the asynchronous methods are primarily of the theme "User's Mental Model", see [17]. Using the laboratory condition many of such problems are also identified. What is typical about these problems are, that the participants' logic is not consistent with the logic of the application. We can also see that the laboratory condition has facilitated the identification of many "information" problems (13), whereas this is not the case for the asynchronous conditions. These problems are mainly concerning lag of information or information that is not understandable by the user.

DISCUSSION

We have evaluated three remote asynchronous methods and compared these to each other and a Lab test as a benchmark. In this section we discuss our findings for the individual methods and examine how our results compare to those from the articles presented under related work.

UCI

The problem reports from this method were the easiest to translate into usability problem descriptions, which corresponds to the findings of [4]. We found that you could almost "copy-paste" many of the reported problems. This is undoubtedly because the participants are forced fill out certain fields and thereby provide specific information. The UCI condition facilitated in the identification of a single cosmetic problem. A similar tendency is shown in [1] as no cosmetic problems are identified by the UCI method. In general the UCI participants found it easier to report usability problems compared to the Diary and Forum participants.

When comparing our results to the results presented in related work we see a clear difference. We have identified 28% of the problems also identified with the laboratory condition. The one that comes closest is [1] with 37% of the problems. On the contrary we find [4] and [27] with

respectively 76% and 60% percent of the problems identified using UCI. This may be due to the difference in training, as we, like [1], have given the participants written instructions. [4] used video training and exercises as well and [27] used an online training tool. When looking at the efforts spent on conducting a UCI test, we see the method as effective, as does [1].

Forum

Using a report form with the possibility for the participants to mutually discuss the problems we hoped to achieve a more nuanced description of each problem. This required actual discussion to take place, which was very limited. The problems that were discussed did however give a clearer understanding of especially what led to a given problem. The participants provided us with documents containing the problem descriptions. In some cases these documents contained a greater number of problems and more detailed problem descriptions than the forum posts. We have also observed that the first post of a thread tends to influence the following posts. If the first post provides little detail, the following posts will so too. The same goes for categorization, although the Forum participants found it easier to categorize the problems than the UCI and Diary participants.

An impartial moderator might have been able to get the discussions going and remind the participants to post categorizations, but we were not aware of these problems before the test had ended, as we did not want to get biased by observing the results before the analysis.

The only article concerning the use of a forum has proven this report form to be successful in identifying usability problems [16]. 3500 people were asked to comment on the application through the forum and about 300 posts were used for identifying problems. We cannot compare the results to ours as no information on number problems identified and resources spent are given.

Diary

The longitudinal aspect of the Diary was hoped to give the participants greater experience in problem identification and reporting and identify unique problems only identifiable during prolonged use of the program. The problem descriptions did not improve over time and the extra four days provided a total of 7 problems, only 4 of these being unique for the Diary condition. The unstructured nature of the diaries required a greater amount of interpretation resulting in a more pronounced evaluator effect. As an example we saw two very different interpretations of the same problem and the text did not indicate what the right interpretation was.

We experienced that 11 problems from this condition was uncategorized. This could be a consequence of the more unstructured form. Another reason for this can be seen in the submitted diaries, where the participants had a tendency to describe and categorize a single large problem, which in fact ought to have been split into multiple minor problems.

A comparison with the results from the article mentioned in related work, that also uses a diary [26], is unfortunately not possible, as there are no results or comments about how well this method performed.

CONCLUSIONS

Our aim in this study was to examine empirically whether or not non usability experts are able to identify and report experienced usability problems and what effects this has on the number of problems identified and the person hours spent.

We found that the test participants were able to identify and report the experienced usability problems in varied, but sufficient, detail. By using the remote asynchronous methods we were able to identify 50% of the critical problems found via the Lab. When using the asynchronous methods we generally found much less serious and cosmetic problems, with the exception of the Diary method, which facilitated in the identification of the same number of cosmetic problems as the Lab. Taking the time spent on conducting the tests and analyzing the results into consideration, the asynchronous methods required much less person hours than the Lab. The fastest of the asynchronous methods was UCI, in which we spent 1/12 of the person hours compared to the Lab.

Limitations

We were not able to control as many variables in the asynchronous settings as in the laboratory setting, e.g. we cannot be sure that all test participants completed all tasks even if stated so or whether the tasks were correctly solved. However, we had the advantage of getting results that were not influenced by an artificial setting.

Future Work

In the future it would be interesting to study other aspects of the asynchronous methods e.g. how training affects the results or the limitations of the methods. A more narrow focus on each of the methods would be interesting as well e.g. examining how to make participants more active during forum discussions.

REFERENCES

1. Andreasen, Morten S. et. al. What Happened to Remote Usability Testing? An Empirical Study of Three Methods.
2. Brush, A.J. Bernheim et. Al. A Comparison of Sunchronous Remote and Local Usability Studies for an Expert Interface. *CHI 2004*.
3. Capra, Miranda G. An Exploration of End-User Critical Incident Classification (Blacksburg, Virginia 2001).
4. Castillo, José C. The User-Reported Critical Incident Method for Remote Usability Evaluation (Blacksburg, Virginia 1997).

5. Castillo, José C. et. al. Remote Usability Evaluation: Can Users Report Their Own Critical Incidents?
6. Desurvire, Heather W. Faster, cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing? *John Wiley & Sons, INC*. Pages 173-202. 1994.
7. Dray, Susan and Siegel, David. Remote Possibilities? International Usability Testing at a Distance. *Interactions*. 2004.
8. Følstad, Asbjørn et. al. Usability Analysis and Evaluation of Mobile ICT Systems.
9. Hammontree, Monty et. al. Remote Usability Testing. *Interactions*. 1994.
10. Hartson, H. Rex and Castillo, José C. Remote Evaluation for Post-Deployment Usability Improvement. *Proceedings of AVI '98* (L'Aquila, Italy 1998), 22-29.
11. Hartson, H. Rex et. al. Remote Evaluation: The Network as an Extension of The Usability Laboratory. *CHI 96*. (1996), 13-18.
12. Morten Hertzum and Niels Ebbe Jacobsen. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*. 15(1). 183-204. 2003
13. Hilbert, David M. and Redmiles, David F. Separating the Wheat from the Chaff in Internet-Mediated User Feedback Expectation-Driven Event Monitoring. *SIGGROUP Bulletin vol. 20 No 1*. (1999).
14. Kjeldskov, Jesper et. al. Instant Data Analysis: Conducting Usability Evaluations in a Day.
15. Marsh, Stephanie L. et. al. Evaluating a Geovisualization Prototype with Two Approaches: Remote Instructional vs. Face-to-Face Exploratory. *Proceedings of the Information Visualization '06*. (2006).
16. Millen, David R. Remote Usability Evaluation: User Participation in the Design of a Web-Based Email Service. *SIGGROUP Bulletin vol. 20 No. 1*. (1999).
17. Nielsen, Cristian Monrad et. al. It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field. *NordiCHI 2006.*
18. Nielsen, Jakob. Finding Usability Problems Through Heuristic Evaluation. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1992.
19. Nielsen, Jakob and Molich, Rolf. Heuristic Evaluation of User Interfaces. *Proceedings of the SIGCHI conference on Human Factors in computing systems: Empowering people*. 1990.
20. Nielsen, Jakob. Usability Inspection Methods. *Conference companion on Human factors in computing systems*. 1994.
21. Olmsted, Erica L. and Gill, Margaret. In-Person Usability Study Compared with Self-Administered Web (Remote – Different Time/Place) Study: Does Mode of Study Produce Similar Results?
22. Petrie, Helen et. al. Remote Usability Evaluations with Disabled People. *CHI 2006 Proceedings.* (Montreal, Canada 2006).
23. Rubin, Jeffrey. Handbook of Usability Testing. *John Wiley & Sons, INC*. 1994.
24. Scholtz, Jean. A Case Study: Developing a Remote, Rapid and Automated Usability Testing Methodology for On-line Books.
25. Scholtz, Jean and Downey, Laura. Methods for Identifying Usability Problems with Web Sites.
26. Steves, Michelle P. et. al. A Comparison of Usage Evaluation and Inspection Methods for Assessing Groupware Usability. *ACM 2001*.
27. Thompson, Jennifer A. Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation (Blackburg, Virginia 1999).
28. Tullis, Tom et. al. An Empirical Comparison of Lab and Remote Usability Testing of Web Sites.
29. Vermeeren, Arnold P.O.S et. Al. Managing the Evaluator Effect in User Testing. *Human Computer Interaction, INTERACT '03*. 2003.
30. Waterson, Sarah et. al. In the Lab and Out in the Wild: Remote Usability Testing for Mobile Devices. *CHI 2002)*. (2002).
31. West, Ryan and Lehman, Katherine R. Automated Summative Usability Studies: An Empirical Evaluation. *CHI 2006 Proceedings*. (Montreal, Canada 2006).
32. Winckler, Marco A. A. et. al. Remote Usability Testing: A Case Study.
33. Winckler, Marco A. A. et. al. Usability Remote Evaluation for WWW. *CHI 2000*.
34. Äijö, Raila and Mantere, Jussi. Are Non-Expert Usability Evaluations Valuable

C. Summary of the Report: "Barriers when Conducting Usability Tests"

Summary of the Report: "Barriers when Conducting Usability Tests"

Introduction

Before examining how the efforts spent conducting usability tests could be reduced, we examined how software development companies in our local area consider usability testing, as we had a hypothesis that several barriers existed that prevented companies from conducting usability tests. We have set up the following hypotheses that we wanted to approve or disapprove:

- Very few companies conduct usability tests.
- The development method influences companies' use of usability tests.
- Mainly large companies conduct usability tests.
- Many different understandings of usability test exist and it is often seen as a test of functionality.
- Companies that conduct usability tests experience problems such as high resource usage (time and money) and limited knowledge about usability testing.
- Companies that conduct usability tests experience advantages such as improved quality, less errors and satisfied customers.
- Companies that do not conduct usability test have prejudices against usability testing such as high resource usage (time and money) and limited knowledge about usability testing.
- The problems experienced by the companies that do conduct usability tests are the same as those that companies that do not test consider as problems.

Method

An electronic questionnaire was sent to 74 software companies of different size in Northern Jutland. The companies were not single person or hobby businesses and they all developed software utilizing a graphical user interface. They had all been contacted in advance and agreed to answer the questionnaire. We got 39 answers. A chi square test shows that when looking at the size of the companies and the products they produce, they are representative for the 74 companies contacted.

In short the companies were asked about:

- General information about the company.
- General information about their products.
- Their development method(s).
- How they understand usability test.
- Whether they conduct usability tests.
- Why they do not conduct usability test.

- What problems and advantages they have experienced when conducting usability tests.
- How usability testing is done.
- General information about the persons conducting usability tests.

Questions were designed as multiple choice and, where appropriate, as open-ended questions. The answers from the open-ended questions were analyzed individually by three analysts using grounded theory [1] to extract the central elements.

Results

In this section the results concerning each hypothesis is presented.

Very few companies conduct usability tests

Before answering whether the companies conduct usability tests or not, we provided a definition of usability testing based on ISO standard 9241-11, and asked them to use this definition while answering the questionnaire. 30 of the 39 companies have answered that they conduct usability tests, either internally in the company, done by an external company or both. This goes against our hypothesis. Some of the answers indicate that not all of the respondents have read the definition thoroughly and we can assume that not all of them conduct usability tests living up to the ISO definition. Considering this it is difficult to say exactly how many of the companies that conduct usability tests.

The development method influences companies' use of usability test

When comparing the companies that conduct usability tests to those that do not, we see a pronounced difference in two places. 3 of the 9 companies that do not conduct usability tests use an undefined development method whereas 3 of the 30 companies that do test use a undefined development method. 1 of the 9 none-testing companies uses an agile development method whereas 8 of the 30 testing companies use an agile method. Companies that do not test have a tendency to use an undefined development method and not use an agile development approach.

Mainly large companies conduct usability test

Size only matters little in this case. Both small and large companies conduct usability tests. 11 of the 30 companies conducting usability tests have more than 11 employees and the same goes for 3 of the 9 companies that do not test.

Many different understandings of usability test exist and it is often seen as a test of functionality.

18 respondents see usability testing as a test of functionality, as our hypothesis suggests. The most common understanding of usability testing among the respondents is; that it is a test involving users or focusing on the users, which 31 have mentioned. Other less common understandings are tests involving task solving, experimenting and focusing on the customer needs.

Companies that conduct usability tests experience problems such as high resource usage (time and money) and limited knowledge about usability testing.

The problems that the companies have had in conjunction with usability testing are summarized here:

- *High resource usage.* This is the most common problem and it is mentioned by 10 of the 30 companies. For some the resource usage is surprisingly high and one respondent mentions a longer time to market.
- *Test participants.* 4 respondents have experienced problems concerning test participants. This involves finding suitable participants and motivating the participants.
- *Developers' way of thinking.* 7 respondents mention that the developers have problems thinking like the users, or they do not take usability testing serious.
- *Implementation.* 3 companies have had difficulties implementing usability testing in their development projects.
- *Customer participation.* Some customers do not see the purpose and need for usability testing, as 5 respondents report.
- *None.* Surprisingly 7 companies have not had any problems conducting usability tests, which might have something to do with the way they define usability testing.

Companies that conduct usability tests experience advantages such as improved quality, less errors and satisfied customers.

The advantages that the companies have experienced in conjunction with usability testing are summed up here:

- *System enhancement.* 17 of the 30 companies have experienced system enhancements in one way or another. In general the answers gotten here are not very specific but cover subjects such as less flaws and better functionality.
- *Customer- and user satisfaction.* 10 companies experience a higher satisfaction among customers and users.
- *Higher sales.* 5 companies have experienced that usability testing of their products differentiates them from competitors, thereby enhancing the sales.
- *Good in conjunction with prototyping.* 3 respondents think that usability testing fits well in development projects using prototyping as development method.
- *New knowledge.* Through usability testing the developers can gain new knowledge about how users see a product and thereby make it easier for them to think like the users, says 5 of the respondents.

Companies that do not conduct usability test have prejudices against usability testing such as high resource usage (time and money) and limited knowledge about usability testing.

The reasons why companies do not conduct usability tests are summed up here:

- *High resources.* 5 of the 9 respondents think that conducting usability tests requires both a lot of time and money. For one company this is because it requires external assistance.
- *Test of functionality has higher priority.* 3 respondents prioritize test of functionality higher.
- *The customer's responsibility.* 2 companies move the responsibility doing usability test to the customer.
- *Unnecessary.* For 3 companies usability testing is seen as unnecessary. One respondent says that the web pages that they develop are so simple that it is not necessary.

The problems experienced by the companies that do conduct usability tests are the same as those that companies that do not test consider as problems.

The companies that do not conduct usability tests are right in the fact that conducting usability tests requires many resources, when we consider the answers from those who do conduct usability tests. This is the only point that they have in common.

Conclusion

Most software development companies among our respondents conduct usability tests in one way or another, which was unexpected. There are, however, differences in the ways that the respondents understand usability testing. Many see it as a test of functionality, but the majority sees it as a test with user focus. The respondents have provided some problems and advantages concerning the use of usability test. These are few compared to the number of respondents. The most common problem is high resource usage and the most common advantages are system enhancement and higher customer and user satisfaction. The companies that do not conduct usability test see high resource usage as the main barrier.

Bibliography

1. Strauss, Anselm and Corbin, Juliet. Basics of Qualitative Research - Techniques and Procedures for Developing Grounded Theory. *SAGE Publications, 2. edition*, 1998.

D. Document Used for Training in the Identification of Usability Problems

2. vejledning-brugervenlighedsproblemer.pdf

Vejledning i beskrivelse af brugervenlighedsproblemer

Formålet med en brugervenlighedstest kan eksempelvis være at finde ud af, hvor godt et program er udformet og opbygget, så det for målgruppen er:

- Let at lære og let at anvende
- Er tilfredsstillende at anvende
- At produktet er funktionelt
- At produktet er tilpasset målgruppen

Et brugervenlighedsproblem er opstået, når du under løsningen af de tilsendte opgaver, oplever episoder, hvor du måske bliver forvirret, hvor du har svært ved at løse en opgave eller slet ikke kan løse opgaven. Ofte opleves det også at man, som bruger, bliver irriteret i lettere eller større grad, over at det ikke umiddelbart er til at gennemskue, hvorledes opgaven kan løses ved brug af systemet. Ofte skyldes brugervenlighedsproblemer, at opbygningen, ikoner eller "ledetekster" i systemet ikke er logiske. Disse ting kendetegner, at der er brugervenlighedsproblemer i systemet.

Kategorisering:

Når du har fundet et brugervenlighedsproblem skal du kategorisere dette, som indrapporteres sammen med problembeskrivelsen. Du skal kategorisere alle dine problemer som enten værende "Lav", "Medium" eller "Høj", alt efter hvor stor gene problemet var for dig under testen.

Det du skal kategorisere er alvorlighedsgraden, se venstre spalte i tabel 1. Når du skal kategorisere, skal du tage udgangspunkt i de udsagn, der er i tabellen. Disse udsagn er kun vejledende, og du behøver f.eks. ikke nødvendigvis at have oplevet irritation, for at et problem kan kategoriseres som værende f.eks. "medium".

Alvorlighedsgrad:	**Forsinkelse i udførsel af opgaven**	**Irritation**	**Afvigelser ift. det forventede**	**Brug for hjælp til løsning af opgaven**
Lav	Mindre end 30 sek. forsinkelse	Blev lettere irriteret	Mindre afvigelser ift. det jeg forventede	Benyttede ikke hjælp eller lidt hjælp til at løse opgaven.
Medium	Mere end 30 sek. forsinkelse	Blev gennemsnitligt irriteret	Betydelige afvigelser ift. det jeg forventede	Benyttede middelgrad af hjælp til at løse opgaven
Høj	Det var ikke muligt at udføre Opgaven	Blev meget irriteret	Store afvigelser ift. til det jeg forventede	Benytte hjælp i stor grad for at løse opgaven

Tabel 1 Kategoriseringer til brugervenlighedsproblemer

Hvis du f.eks. har oplevet et problem:

- Som forsinkede dig mere end 30 sekunder i at udføre opgaven, fx fordi du ikke kunne finde den menu, som du skulle bruge for at kunne udføre opgaven,
- Afveg opbygningen, ”ledeteksterne” eller ikonerne af/i systemet betydeligt fra det, som du forventede

Ja, så kan det oplevede brugervenlighedsproblem kategoriseres som ”Medium”.

Eksempler på brugervenlighedsproblemer og deres kategorisering.

Nedenstående er eksempler på brugervenlighedsproblemer.

Eksempel 1:
En bruger benytter Microsoft Word og vil gerne indsætte en tekstlinie øverst på dokumentet, som går igen på alle sider. Brugeren ved godt, at det er noget med at *indsætte* et sidehoved, og åbner derfor menuen ”Indsæt”, gennemser mulighederne, men der er ikke noget som er kaldt ”sidehoved”. Brugeren prøver så derefter at finde funktionen i menuerne ”Funktioner” og ”Formater”, for det kan jo tænkes, at funktionen ligge der, men nej. Brugeren føler sig nu lettere irriteret over, at det ikke er muligt at finde funktionen.

Brugeren vælger nu at benytte hjælpefunktionen i Word og vælger menuen ”Hjælp”, skriver sidehoved i søgefeltet og trykker ”Søg”. Her kommer et link frem, hvor der står ”Indsætte sidehoveder eller sidefødder”. Dette vælges af brugeren, og beskrivelsen til, hvorledes det er muligt at indsætte et sidehoved, kommer frem. Det viser sig, ud fra beskrivelsen, at funktionen ligger i menuen ”Vis” – ikke særligt logisk, tænker brugeren og bliver lettere irriteret over den tåbelige opbygning. Brugeren kom dog forholdsvis hurtigt over problemet ved brug af hjælp.

Ved et kig på tabel 1, kan man se, at dette problem kan kategoriseres til ”Lav”, da brugeren ret hurtigt fandt ud af at løse problemet vha. hjælpefunktionen, og brugeren kun kortvarigt blev opholdt af problemet (under 30 sek.). Systemets opbygning var ikke helt som forventet og brugeren blev lettere irriteret.

Eksempel 2:
En bruger vil på Told og Skats hjemmeside søge efter, hvor stor en del af det betalte børnebidrag (til børn fra tidligere ægteskab), der er fradragsberettiget. Det er ikke det fulde beløb, som er fradragsberettiget.

Brugeren finder Told og Skats hjemmeside, www.toldogskat.dk, hvilket går fint, her vælges linket ”Borger”, eftersom han jo er borger og spørgsmålet ikke har noget med en virksomhed at gøre, hvilket er en anden valgmulighed.

Brugerens tanke er, at det letteste må være at søge efter ordet ”børnebidrag” i søgefeltet, så dette gør brugeren. Men alle de mulige links som kommer frem, er enten hvilken kolonne på selvangivelsen, der skal anvendes til børnebidrag, nogle tekniske detaljer i ligningsloven eller skatteretssager, som har været afholdt omkring emnet. Taksten for fradraget, som var den ønskede oplysning, er ikke at finde. Brugeren må til

sidst give op efter at have prøvet adskillige af de links, der var til rådighed under emnet. Brugeren må så i stedet ringe til Told og Skat dagen efter (for nu er der lukket), hvor der for øvrigt ofte er lang ventetid på telefonen – øv, tænker brugeren, hvad f…. har de en hjemmeside for, når man ikke kan finde de oplysninger, man skal bruge!.

Hvis vi igen ser i tabel 1, kan dette problem kategoriseres til ”Høj”, da brugeren ikke kan få løst sit problem på hjemmesiden. Brugeren bliver stærkt irriteret, da denne ikke finder den ønskede oplysning, til trods for at den hjælp, der er til rådighed, blev benyttet (søgefunktionen).

N.B. Det kan oplyses til denne case, at Told og Skat nu har lagt de ønskede oplysninger på deres hjemmeside, som let kan fremsøges vha. af søgefunktionen. Takster er her dateret i efteråret 2006.

Eksempel på rapportering af problemer

Herunder ses et eksempel på hvordan skemaet til indrapportering af problemer kan udfyldes. Eksemplet stammer fra eksempel 1 ovenfor, som omhandler et problem med Microsoft Word.

Angiv opgavens nummer:		Opgave 1
Angiv navnet på det vindue i Thunderbird, hvor problemet opstod:	Eksempel	Microsoft Word (hovedvindue)
Beskriv hvad du forsøgte at gøre, da problemet opstod (din intention):		Jeg ville indsætte et sidehoved i mit dokument.
Beskriv hvad du troede systemet ville gøre, da problemet opstod (din forventning):		Jeg havde en forventning om at funktionen til at indsætte sidehoved lå i menuen "indsæt".
Med så mange detaljer som muligt, beskriv da problemet og hvorfor du tror, det opstod:		Det var ikke muligt for mig at finde funktionen til at indsætte sidehoved i menuen "indsæt". Jeg forsøgte også at lede i menuerne "formater" og "funktioner", dog uden held. Det virkede en smule frustrerende.
Var det muligt for dig at omgå problemet?		Ja
Beskriv hvordan du har forsøgt at omgå problemet (eller hvordan du har løst dette):		Jeg søgte efter "sidehoved" med hjælpefunktionen og fandt frem til at funktionen lå i menuen "vis". I hjælp kaldes det desuden at "Indsætte sidehoveder og sidefødder", selvom funktionen ikke ligger i menuen "indsæt", hvilket er lidt forvirrende.
Kan du reproducere problemet og få det til at opstå igen?		Ja
Vurder alvorlighedsgraden af problemet?		Lav alvorlighedsgrad

E. Summary

Summary

Our motivation behind this master thesis comes from the fact that it is well known that the conduction of usability tests in a laboratory setting and especially the following video analysis is very resource demanding and expensive. Another problem regarding this form of usability testing is the cost of bringing test participants to the laboratory, especially when these are situated in another country. Considering the increasing global scope of software development, e.g. outsourcing and development of software for use in foreign countries, this problem will most likely increase in the future.

To overcome these barriers it is possible to apply various discount methods. However, the most widely applied "discount" usability methods do not involve users. Heuristic inspection is one of these types of discount usability methods. Here usability experts inspect a given system for usability problems using a set of heuristics.

In this master thesis we have examined how it is possible to reduce the resources spent on conducting user based usability tests, analyzing the results and acquiring test participants.

To answer our questions we have made two empirical studies:

- In the first study we have evaluated a user based discount method called Instant Data Analysis (IDA). We conducted a user-based think aloud laboratory usability test and analysed the test results using two different methods, a traditional Video Data Analysis (VDA) and IDA. The IDA method is used for analysing the results from a user based laboratory test, with the aim of quickly identifying the most severe usability problems. The results from using the two methods were compared and analysed.

- In the second study we compared three different methods for remote asynchronous usability testing. We conducted a laboratory usability test and three remote asynchronous tests. In asynchronous remote methods the user and the test monitor is separated in time and space. The test participants can, for instance, sit at home, work or wherever possible and participate in the test. The participants' had to identify, describe and categorize the experienced usability problems. The participants reported to us in three different conditions: User-reported Critical Incident (UCI), a Forum and a longitudinal based Diary. The results from using the methods were compared and analysed. These methods save the evaluators from spending time conducting the test in a laboratory as well as using a long time analysing the data. A further benefit is that the test participants and the test monitor can be separated over physical boundaries, hereby reducing the resources required to bring participants to a laboratory.

The key findings from the first study show that by using IDA we were able to considerably reduce the time spent on identification of usability problems. We were able to reveal 68% of the total number of problems using IDA and we found 81% of all problems using VDA. The aim of IDA is to assist in identifying the most severe usability problems in less time, and we found more critical problems using IDA (89%) than we did using VDA (72%). Considering the serious problems we found 76% using IDA and 76% using VDA. We found that IDA required 11.5 person hours and VDA 39.25 person hours. IDA thus fulfills its purpose of revealing the most severe problems in less time than a conventional video data analysis.

In the second study we found that the test participants were able to describe the experienced usability problems in varied, but sufficient, detail. By using the remote asynchronous methods we were able to identify 50% of the critical problems found via the Lab. When using the asynchronous methods we generally found much less serious and cosmetic problems, with the exception of the Diary method, which facilitated in the identification of the same number of cosmetic problems as the Lab. Taking the time spent on conducting the tests and analyzing the results into consideration, the asynchronous methods required much less person hours than the Lab. The fastest of the asynchronous methods was UCI, in which we spent 1/12 of the person hours compared to the Lab.