Let your users do the testing: A comparison of three remote asynchronous usability testing methods

Conference Paper · April 2009

DOI: 10.1145/1518701.1518948 · Source: DBLP

Anders Bruun
Aalborg University

Let Your Users Do the Testing: A Comparison of Three Remote Asynchronous Usability Testing Methods

Anders Bruun[1], **Peter Gull**[2], **Lene Hofmeister**[3], **Jan Stage**[4]
[1]Mjølner Informatics A/S, Finlandsgade 10, DK-8200 Århus N, Denmark
[2]Jyske Bank A/S, Vestergade 8-16, DK-8600 Silkeborg, Denmark
[3]Nykredit A/S, Fredrik Bajers Vej 1, DK-9220 Aalborg East, Denmark
[4]Aalborg University, Department of Computer Science, DK-9220 Aalborg East, Denmark
ab@mjolner.dk, pg@jyskebank.dk, leho@nykredit.dk, jans@cs.aau.dk

ABSTRACT

Remote asynchronous usability testing is characterized by both a spatial and temporal separation of users and evaluators. This has the potential both to reduce practical problems with securing user attendance and to allow direct involvement of users in usability testing. In this paper, we report from an empirical study where we systematically compared three methods for remote asynchronous usability testing: user-reported critical incidents, forum-based online reporting and discussion, and diary-based longitudinal user reporting. In addition, conventional laboratory-based think-aloud testing was included as a benchmark for the remote methods. The results show that each remote asynchronous method supports identification of a considerable number of usability problems. Although this is only about half of the problems identified with the conventional method, it requires significantly less time. This makes remote asynchronous methods an appealing possibility for usability testing in many software projects.

Author Keywords
Remote testing, asynchronous testing, usability testing, empirical study.

ACM Classification Keywords
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology, Theory and methods.*

INTRODUCTION

User-based testing has become almost a de facto standard in usability engineering. The essential idea is to base assessment of the usability of a system on observation of users working with the system. The classical approach was to combine this idea of user-based testing with the think-aloud protocol and implement it in a laboratory setting, e.g. [27]. A key drawback of this approach was that it turned out to demand considerable resources for planning the tests, establishing a test setting and conducting the tests [8, 9, 22, 23, 24]. A subsequent analysis to identify usability problems, conducted as a rigorous walk-through of the hours of video recordings that documented the users' interaction with the system, was almost equally demanding in terms of resources [17].

The classical approach to usability testing has limited influence on contemporary practice in software development and commercial usability testing. Practitioners and researchers have developed a rich inventory of techniques for usability testing that reduce the resource demands, either for the whole evaluation process, e.g. [22, 23, 24], or for selected activities, e.g. [17]. Modern software organizations combine such techniques with new development approaches, e.g. [2]. At the other end of the spectrum, there are many software organizations that still have no systematic usability activities deployed in their development processes, and the most frequent cause given is perceived resource demands [3].

Software organizations that develop and evaluate products for global markets or practice outsourcing or global software development face different but equally significant obstacles. When developers, evaluators and users are distributed across different organizations, countries and time zones, user-based usability testing, in particular planning and setting up the test, becomes a nearly insurmountable logistic challenge.

The difficulties with usability testing that are emphasized above have led some researchers to inquire into remote usability testing. Remote usability testing denotes a situation where "the evaluators are separated in space and/or time from the users" [7]. There is a basic distinction between remote synchronous and remote asynchronous methods.

With a remote synchronous method, the test users and evaluators are separated in space [6]. The evaluators can observe the users over a network connection by means of video capture software where, for example, the content of the test participants' screen is transmitted to the evaluators who are residing in a remote location [1, 2, 6, 9, 12, 14]. A remote synchronous method still requires the evaluators to be present in real time to control and monitor the test, and the results need to be analyzed by the evaluators. Thus, this method is almost as resource demanding as a typical user-based test, but it escapes many of the logistic problems [9].

With a remote asynchronous method, the test users and evaluators are separated both in space *and* time [6]. This implies that the evaluators no longer need to be present at the time when the users are working with the system. Thereby, the asynchronous methods eliminate a key drawback of the synchronous methods. Moreover, some of the asynchronous methods involve the users in the reporting of usability problems with the aim of also eliminating a detailed analysis conducted by the evaluators.

In this paper, we present results from an empirical study of three methods for remote asynchronous usability testing. In the following section, we describe previous research on remote asynchronous usability testing. Based on this, we have selected three asynchronous methods. Next, we describe the experimental method of the empirical study we conducted to compare the qualities of the three methods. This is followed by a presentation of the results of the study. Here, we focus on the number of usability problems identified and the time spent conducting the usability tests. Then we compare with results obtained by others and discuss our results. Finally, we provide the conclusion.

RELATED WORK

We have conducted a systematic study of literature on remote asynchronous usability testing. In this section, we provide an overview of papers that present empirical studies of the use of one or more remote asynchronous methods for usability testing. Thus papers that only mention or briefly outline remote asynchronous usability testing are not included. Table 1 provides an overview of the 22 papers.

Basis for comparison	**Paper #**
Conventional laboratory	0, 6, 19, 25, 26, 31, 32, 34, 35, 37, 38
Usability expert inspection	5, 6, 7, 13, 14, 29, 30, 38
No comparison	10, 16, 20, 28, 36

Table 1. Empirical papers with remote asynchronous usability testing methods and the comparisons they make.

There are 17 out of the 22 papers that compare remote asynchronous usability testing with established approaches such as a conventional user-based laboratory test, an expert-based usability inspection or both. Out of these, only one compares multiple asynchronous methods [6]. This is, however, only a cost-benefit comparison graph based on intuition and not empirical data. The other 16 papers only deal with a single remote asynchronous method. The remaining 5 papers are documenting empirical studies that apply an asynchronous method, but without comparing it to other approaches.

Table 2 provides an overview of the methods that are used in the papers listed in Table 1.

Method	**Paper #**
Auto logging	20, 28, 29, 30, 34, 35, 36, 37
Interview	10, 26, 29, 30, 35
Questionnaires	10, 19, 25, 28, 29, 30, 32, 34, 35, 36, 37
User-reported critical incident	0, 5, 6, 7, 13, 14, 31
Unstructured problem reporting	10, 19, 38
Forum	21
Diary	31

Table 2. Methods for remote asynchronous usability testing.

Auto logging is a method where quantitative data like visited URL history and the time used to complete tasks are collected in log files that are analysed. This method can show if the paths to complete tasks are well designed [20, 28, 29, 30, 34, 35, 36, 37], but it is lacking the ability to collect qualitative data needed to address usability issues beyond the likes of path finding and time used. Therefore, it is often combined with interviews and/or questionnaires as follow-up. With this combination, auto logging has found many of the same problems as heuristic inspection [30]. In another study, the evaluators used this approach to identify 60% of the problems found via a heuristic inspection [29]. In a comparison with a conventional laboratory method, it is concluded that the auto logging method in comparison is not too efficient [37]. The reports of two of these studies [30, 37] do not provide any information about the total number of usability problems identified with auto logging method. In another comparison, the evaluators using the auto logging method identified 40% of the usability problems found in a conventional laboratory test [34].

The *User-reported Critical Incident* method (UCI) is based on the idea that the users themselves report the problems they experience. This should relieve evaluators from conducting tests and analysing results. It has been concluded that test participants are able to report their own critical incidents, e.g. Castillo shows that a minimalist approach works well for training participants in identifying critical incidents [6]. The first studies of this method concluded that the participants were able to categorize the incidents [6, 7], but a more recent and systematic study concludes that the participants cannot categorize the severity of the critical incidents they identify [1]. The reason for this discrepancy may be that the training conducted by [5, 6, 7] was more elaborate than [1], but the training was done with the researchers physically present [6, 7, 31] which contradicts the idea of remote testing. The training conducted by [1] was done over the Internet. The number of usability problems identified also varies between the different studies. In one of the first studies, 24 participants identified 76% of the usability problems found by experts [6]. In a later study, 10 participants identified

60% of the problems found in a conventional laboratory test [31]. In the most recent study, 6 participants identified 37% of the usability problems found in a conventional laboratory test [1].

Unstructured problem reporting is based on the idea that participants make notes on the usability problems they encounter while working on a set of tasks [10, 19, 38]. The descriptions of the studies of this method are brief and there is no information about the predetermined content they have wanted participants to write down. In a study, 9 participants using this kind of reporting identified 66% of the usability problems found by experts [38]. These researchers recommend a more structured approach which is close to the user-reported critical incident method. Another study showed that 8 participants using the unstructured approach identified 50% of the total usability "issues" found in a conventional laboratory test, but this was based on a procedure where the participants in the remote asynchronous condition solved instructional tasks whereas the participants in the laboratory test solved exploratory tasks. As noted by the authors, this makes the comparison "unfair" [19].

The *forum* has been used as a source for collecting qualitative data in a study of auto logging [20]. The researchers did not specifically encourage the participants to report usability issues through the forum, but the participants did report detailed usability feedback. There is no information about user training or the number of usability problems reported in the forum [20]. In a different study, the author argues that participants may be more motivated to report problems, if reporting is a collaborative effort amongst participants. The author believes that participants through collaboration may give input which increases data quality and richness compared to the user-reported critical incident method [31].

The *diary* has been used on a longitudinal basis for participants in a study of auto logging to provide qualitative information [30]. There is no information about the usefulness of the method or the experiences with it. However, it is mentioned that the participants used their diary on a longitudinal basis to report on the usability problems they experienced with the use of a particular hardware product.

METHOD

We have conducted an empirical study of remote asynchronous usability testing with the following four conditions:

- Conventional user-based laboratory test (Lab)
- User-reported critical incident (UCI)
- Forum-based online reporting and discussion (Forum)
- Diary-based longitudinal user reporting (Diary)

The conventional laboratory test was included to serve as a benchmark. In the rest of this section, we describe the method of the study.

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 Aalborg University, at different faculties and departments. They were between 20 and 30 years of age. They signed up voluntarily after we submitted an email call for participants. Half of them were taking a non-technical education (NT), and the other half a technical education (T). For all four 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. They received no course credit or payment. After completion, we gave each a bottle of wine for their effort.

Training. The test subjects participating in the remote asynchronous sessions were trained in identification and categorisation of usability problems. This was done using a minimalist approach that was strictly remote and asynchronous, as they received written instructions via email, explaining through descriptions and examples what a usability problem is and how it is identified and categorised. The categories were "low", "medium" and "high", corresponding to the traditional cosmetic, serious and critical severity ratings [1].

System. In order to facilitate comparison of the results, we chose to use the same system and tasks as another study that involved both synchronous and asynchronous methods [1]. Accordingly, we tested the email client Mozilla Thunderbird version 1.5. None of the test subjects had used Mozilla Thunderbird before.

Tasks. All participants were asked to solve the following tasks (the same as the ones used by [1]):

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 had fixed tasks across the four conditions to ensure that all participants used the same parts of the system.

Laboratory Testing (Lab)

Setting. The laboratory test was conducted in a state-of-the-art usability laboratory. In the test room, 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 participant was thinking aloud.

Procedure. The procedure followed the guidelines of [27]. It was not conducted by the authors of this paper. The test subjects were introduced to the test sequence and the concept of thinking aloud by the test monitor. We scheduled one hour per participant including post-test interview and switching participants. The interviews were carried out by the test monitor. Participants had to solve the nine tasks while thinking aloud. The test monitor had a timeframe for the completion of each task. Participants who had not solved a task in this time received help from the test monitor to ensure that all tasks were completed.

Data collection. A video of the test subjects' desktop was recorded along with video showing the test subject's face.

This condition was referred to in the introduction as the classical approach. We decided to use it as a benchmark, despite its limited influence on contemporary usability testing practice, because it facilitates comparison with other studies, where it is commonly used. The condition involved thinking aloud, which in discussed in [10]. Again, we used it to facilitate comparison with other studies.

The Three Remote Conditions

Some methodological aspects were common to all three remote conditions.

Setting. In all remote asynchronous methods, participants worked at home using their own computer. The participants could carry out the tasks whenever they wanted; but had to completed by a specified date. Once they started, they had to finish all tasks in one session.

Procedure. We sent all participants the training material and a guide on how to install the system. The participants were asked first to go through the training material, install Mozilla Thunderbird and then begin task completion. With each task there was a hint that allowed the participant to check whether they had solved the task correctly. In the Lab condition, you can control the users' task solving process in accordance with the correctness of their solution. Remote users need a similar criterion in order to know when they can stop the work on a task. That is the purpose of the hint.

User-Reported Critical Incident Method (UCI)

Procedure. 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 that was programmed using PHP, JavaScript and an MySql database. The content of the form was similar to that used by [5] and [31]. 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 log of time spent completing each task and email this log to us.

Data collection. At the bottom of the online form was a submit button. When it was pressed, the data was saved in an online database and the form was reset, ready to enter a new entry. The form was running in a separate browser window, requiring the participants to toggle between windows when they encountered a problem. Reporting might be integrated directly into the application [13], but it requires extra resources to implement, and the two-window approach has been shown to work as well [31].

Forum

Procedure. After installing the system, the participants were asked to first take notes on the usability problems they experienced during completion of the tasks and also to rate the severity. The participants were asked to finish all tasks in one sequence and to create a log of the time taken to finish each task. After completion of the tasks the test participants were instructed to uninstall Mozilla Thunderbird to avoid confounding longitudinal use. They were then asked to post and discuss their experienced usability problems with the other participants. They were given a week for that. 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, comment on this either by posting an agreement with the problem description and categorization or state a disagreement with a reason.

Data collection. The forum in itself is a data collection tool, so the data collection for this method was very simple.

Diary

Procedure. The participants were given a timeframe of five days to write about experienced usability problems and severity categorizations in their diary. We provided the same list of elements to consider as to 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 impose any formal content structure. On the first day, the participants received the same nine tasks as all other participants. We instructed them to complete those nine tasks on the first day, and then email the experienced problems and the log to us immediately after completion.

During the remaining four days the participants received additional tasks to complete on a daily basis. These tasks were of the same types as the original nine tasks. The purpose was to generate longitudinal data by ensuring that the users kept working with the system through all five days and making daily entries in their diary. The longitudinal element was only introduced after the first day to ensure that the data from that day were comparable with the other conditions and the value of more days could be identified.

Data collection. The participants e-mailed their diary notes to us after the first and the fifth day.

Data Analysis

The data analysis was conducted by three of the authors of this paper. Each analyzed all data from all four test conditions. This consisted of 40 data sets, 10 for each condition. All data was collected before conducting the analysis. Each data set was given a unique identifier, and a random list was generated for each evaluator, defining the order of analysis of all data sets. Each evaluator analyzed all the data sets alone, one at a time.

For the Lab condition, the videos were thoroughly walked through. The data from the three remote conditions was read one problem at a time. By using only the information available in the users' problem description, it was transformed into a usability problem description. If necessary, Mozilla Thunderbird was checked to get a better understanding of the problem. When analyzing forum descriptions, previous problem descriptions by other users in the same forum thread was also included in the analysis. If a description could not be translated into a meaningful problem in short time or we could not identify the problem using Thunderbird, the problem was not included in the problem list, because we wanted to make sure it reflected the users' experience and not problems made up by the evaluators. There were 12 user descriptions (1 from UCI, 3 from Forum and 8 from Diary) that could not be translated into a problem description because the description was impossible to understand or missing.

During the analysis we also rated the severity of the problems, as we wanted to make sure that this was done consistently.

When each evaluator had created a problem list for all data sets, they merged their lists for each of the four conditions. These four lists were then merged to form a complete problem list for the individual evaluator. The three evaluators then merged their individual lists for each of the four conditions in order to create a joint problem list for that condition. In case of disagreement, they negotiated by referring to the system and the original data until they reached an agreement. Severity rating in the joined lists was done using the most serious categorization. The joined lists for each condition were then joined to form a complete joined problem list. The resulting problem list included a detailed description of each usability problem.

	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.

Hertzum and Jacobsen [15] have shown that evaluators do not find exactly the same usability problems from the same data set. They call this the evaluator effect. To verify the agreement between evaluators, the evaluator effect can be calculated as an any-two agreement showing to what extent the evaluators have identified the same problems [15]. Table 3 shows the average any-two agreement for all of the test conditions and for the entire test. Compared to Hertzum and Jacobsen's findings [15], our any-two agreement is very high.

RESULTS

This section presents the findings from the study.

	Lab		**UCI**		**Forum**		**Diary**	
	N=10		N=10		N=10		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. Percentages are of the total number of problems shown in brackets in the left column.

Number of Problems Identified and Time Spent

In this section we compare the four conditions with respect to the number of usability problems identified and the time spent on analysis. There is considerable variation in the standard deviations between the conditions. As we have no data on the task solving process in the remote conditions, we cannot explain this variation. An overview of the problems identified can be seen in table 4.

Table 5 shows the time spent in the four conditions. All time indications are the sum for all evaluators involved in that activity. The time for preparation does not include time spent on finding test subjects, as this was done jointly for all four conditions. In total, it took about 8 hours. Task specifications were taken from an earlier study. Preparation in the three remote conditions was primarily to work out written instructions for the participants. These instructions could to a large extent be reused between the conditions, thus a test with only a single remote method would require a few hours more.

	Lab (46)	UCI (13)	Forum (15)	Diary (29)
Preparation	6:00	2:40	2:40	2:40
Conducting test	10:00	1:00	1:00	1:30
Analysis	33:18	2:52	3:56	9:38
Merging problem lists	11:45	1:41	1:42	4:58
Total time spent	61:03	8:13	9:18	18:46
Avg. time per problem	1:20	0:38	0:37	0:39

Table 5. Person hours spent on test activities. The numbers in parentheses are the total number of problems identified.

The UCI condition included setting up a web based form that the participants should use to report each problem they identified. This was made by first developing a tool and then using this tool to create the form. This took about 16 hours, but the tool and most of the form is directly reusable for a new test, which is why we have excluded this time.

The time for conducting the test in the Lab condition is the time spent by the test monitor. This includes the time it took the user to solve the tasks as well as setting up the system, briefing the user and administering a questionnaire.

	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=0.01423 *
Diary	p=0.0031 **	p=0.002 **	p=0.01423 *	

Table 6. Fisher's exact test for the total number of usability problems identified in the four conditions. * = significant difference, ** = Very significant difference, * = Extremely significant difference**

Lab

From the Lab test we identified a total of 46 usability problems. Twenty 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 reported problems using the Lab condition. As many as 95% of the critical problems, 82% of the serious and 50% of all cosmetic problems were found using this method. Thus the Lab condition identified more problems than the others, but at the same time it was the most time consuming as we spent 61 hours on it.

Lab vs. UCI

The UCI condition revealed 13 problems, consisting of 10 critical, 2 serious and 1 cosmetic. A Fisher's exact test gives an extremely significant difference (see table 6 for an overview) meaning that the Lab condition identified more problems than UCI. Fisher's exact test for each level of severity gives p=0.00139 for critical problems, p<0.001 for serious and p<0.001 for cosmetic. Thus for each severity level, the Lab condition finds more problems than UCI.

There was some overlap between the problems identified in the two conditions. All the 10 critical problems found with UCI were also found with the Lab condition. One of the 2 serious problems found with UCI was also found in the Lab condition. The single cosmetic problem found with UCI was not found in the Lab condition.

The UCI condition was clearly less time-consuming as we only spent just over 8 hours compared to the 61 hours for the Lab condition.

Lab vs. Forum

With the Forum condition we could identify a total of 15 problems (9 critical, 1 serious and 5 cosmetic). A Fisher's exact test reveals an extremely significant difference (p<0.001), meaning that the Lab condition identified significantly more problems than the Forum condition. For the individual severity levels, Fisher's exact test gives p<0.001 for the critical, p<0.001 for the serious and p=0.0687 for the cosmetic problems. Thus the Lab condition is significantly better for the critical and serious problems, but no significant difference for cosmetic ones.

In the Lab condition we found all the 9 critical problems that were identified by the Forum. The serious problem from the Forum condition was also identified in the Lab condition, and 3 of the 5 cosmetic problems were also in common between the Lab and Forum.

We spent just over 9 hours on the Forum condition compared to the 61 hours on the Lab condition.

Lab vs. Diary

The Diary condition revealed a total of 29 problems, consisting of 11 critical, 6 serious and 12 cosmetic. A Fisher's exact test shows that the Lab condition identified significantly more problems than the Diary condition. For the severity levels, the Lab condition found significantly more critical (p=0.0036) and serious problems (p=0.013), whereas there was no significant difference for the cosmetic problems (p=1.00).

Out of the 11 critical problems found with the diary condition, 9 were also revealed by the Lab condition. For the serious problems, 3 of the 6 Diary problems were also found in the Lab condition. Finally, 3 of the 12 cosmetic problems were also found with the Lab method.

The time spent on the Diary condition was just under 19 hours compared to the 61 hours in the Lab condition.

UCI vs. Forum

The UCI and Forum conditions have 5 critical problems in common and did not find the same serious or cosmetic problems. A Fisher's exact test reveals no significant difference (p=0.6639) in the total number of problems identified between the two methods. Thus we cannot identify any one of them as performing best overall. There is no significant difference for the severity levels either.

We spent almost the same time on the UCI condition, just over 8 hours, and Forum conditions, just over 9 hours.

UCI vs. Diary

Using the UCI and Diary conditions we found 7 critical problems and 2 serious problems common for both methods. On the total number of problems identified we found a very significant difference (p=0.002), meaning that the Diary condition performed better than UCI. For the individual severity levels, the only significant difference was on cosmetic problems, where the Diary condition found more problems (p<0.001).

We spent just under 19 hours on the Diary condition and just over 8 hours on the UCI condition.

Forum vs. Diary

With the Forum and Diary conditions we found 7 critical problems, 1 serious and 1 cosmetic problem common for both methods. The difference on the total number of problems identified is significant (p=0.0142), meaning that the Diary condition performed better than the Forum. For the three severity levels, there is no significant difference.

We spent just under 19 hours on the Diary condition compared to just over 9 hours on the Forum condition.

Task Completion

For all 40 test participants the mean value of completed tasks is 8.9 out of 9. The only condition, in which not all tasks were completed, was UCI, where one test participant did not complete tasks 3 and 4, because that person had no understanding of a filter and how it worked. All participants completed the 9 tasks in the Lab condition, but the majority of participants experienced difficulties in completing tasks 3, 6 and 7. For task 3 and 7, it was caused by limited or no understanding of filters, and for task 6, it was a difference compared to Outlook that confused some users.

Task Completion Time

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

The standard deviation emphasizes 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

Some of the problems we have identified were only found in one condition. Table 7 gives an overview of the number of problems identified in one test condition only.

	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 in the Lab condition. The majority of these are serious and 24% of all critical problems identified are only identified in the laboratory condition. Looking at all three severity categories, the unique Lab problems are primarily within the theme "Information", cf. [21], i.e. problems where the participants were missing information or did not understand the information from the system as it was too technical. The UCI and Forum conditions, also being the ones revealing the smallest number of problems in total, have revealed 3 unique problems in total, none of them being critical. The diary has revealed even more unique cosmetic problems than the laboratory condition (9). The unique problems found via the Diary condition are distributed evenly over the different problem themes defined in [21].

The Lab condition identified 5 critical problems not found by any of the remote methods. Interestingly, this means that by combining the results from the UCI, Forum and Diary conditions, we have identified 16 of the total of 21 critical problems. The total time for all three remote conditions sums up to about 36 hours, which is just over half of the time spent on the Lab condition.

Differences between Severity Ratings

All the users in the three remote conditions received the same instructions on rating of the severity of usability problems. We explained it to the users as categories, but they translate directly into a standard three-level severity rating scale. In this section, we examine whether it was possible for the participants from the remote conditions to categorise the problems properly. Table 8 shows how the participants' severity ratings corresponded to the ones made 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 8. 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 where the severity ratings best matched the ones made by the evaluators. In this case, the ratings of only 3 of the 13 problems (23%) did not match, and all problems were rated. The reason for this was that it was not possible to report a problem without a severity rating. In the Diary condition, as many as 11 out of 29 (38%) problems were not rated. The Forum method is, like the Diary, a more unstructured approach than UCI. In this condition, 5 problems out of 15 (34%) were not categorized similarly, including problems without a rating.

Differences 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", cf. [21]. In the Lab condition many such problems are also identified. What is typical about these problems are, that the participants' logic is not consistent with the logic of the system. We can also see that the Lab condition has facilitated the identification of many "information" problems (13), as opposed to the asynchronous conditions. These problems are mainly concerning lack of information or information that is not understandable by the user and confirm the value of the think-aloud protocol in the Lab condition.

DISCUSSION

With the three remote asynchronous methods we have studied, detailed analysis of usability problems is replaced with an activity where descriptions made by the users are transformed into a list of usability problems. With UCI, this transformation was very simple. You could almost "copy-paste" several of the users' descriptions directly into the problem list. This also corresponds to the findings of others [6]. The reason is undoubtedly that the participants were forced to fill out certain fields and thereby provide specific information. The output from the forum also required very little work, but the quality differed considerably. The idea of the forum was to allow participants to discuss the problems collaboratively. Thereby, we hoped to achieve a richer 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 longitudinal aspect of the Diary condition was intended to give the participants a better basis for problem identification and reporting and enable them to identify problems that were only identifiable during longer use of the program. The problem descriptions did not improve over time and the extra four days only 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 and considerably more time spent on analysis.

The research literature on remote asynchronous usability testing is limited. Therefore, we have only few possibilities for comparison our results with others. The benefits of the UCI method have been identified before [6]. With this method, we identified only one cosmetic problem. A similar tendency is reported in [1]. On the number of problem identified with UCI, there are some remarkable differences. We found 28% of the problems found in the laboratory condition. This is close to a previous result of 37% [1], but very different from earlier results of 76% [6] and 60% [31]. This may be due to the difference in training, as we, like [1], have given the participants written instructions. [6] used video training and exercises as well, and [31] used an online training tool.

The only reference on the Forum method has no information about the number of problems identified and the resources spent [20]. In a recent study [18], 2 forums were used to evaluate the user experience with two different versions of the same game. The two forums worked out differently, with one producing more relevant and detailed information than the other. The reason seemed to be that they started out differently. Based on that study and the one reported in this paper, we would suggest the Forum method should be extended with a moderator to ensure a good start, enough details in the descriptions and ratings of severity.

Research on the Diary method is very limited. The single reference [30] we have found has no results or comments about how well the method performed. In our study, the Diary condition required more time for analysis. This could, however, be reduced if the diary was combined with the problem reporting format that is used with UCI, but that would change the whole idea of the diary.

Another difference between the Lab and the three remote asynchronous conditions is the training of users. We received very few useless descriptions from our users, but as emphasized above the number of problems identified was lower than those reported from some of the other studies. It would be interesting to experiment with the effect of different training formats and materials.

One of the difficulties in our study was that we did not observe the participants in the remote conditions. This is in line with the philosophy of the methods, but the consequence is that we have missed information about their task solving process. It also means that the task completion times have to be read with great caution.

The number of usability problems identified is a key element in our comparison of the different conditions. This is typical in the research literature on method comparisons. It would be relevant to compare based on other measures, e.g. the method's ability to reveal the users' understanding of the system or the usefulness of the problems identified.

The Lab condition has been used as a benchmark, although it has limited influence in practice. The reason is to facilitate comparison with related work. It would be interesting to conduct a follow-up study where the remote conditions are compared to more modern approaches.

Our study was based on a system that was a finished product. A main challenge for usability engineering is to conduct testing early in the development process. This may be more difficult with a remote asynchronous method, because the users are on their own with no possibility of getting support with a system that is not fully functional.

CONCLUSIONS

This paper has reported from an empirical study of three remote asynchronous usability testing methods. The methods were compared to each other and to a classical laboratory-based approach. On the overall level, the three remote methods performed significantly below the classical lab test in terms of the number of usability problems identified. For critical and serious problems, the Diary condition, which was the best of the remote methods, identified only half of the problems found in the Lab condition. The other two remote methods performed similarly for critical problems but worse for serious problems. This may seem disappointing. Yet two of the remote methods produced these results with an effort that only amounted to about 13% of what the lab test took. The diary method took more time but still only about 30%.

This makes the remote methods an appealing possibility for many software projects. It is often highly relevant to get a cheap usability test although it is not complete. In that case, one of the remote tests would be an interesting possibility. In addition, the remote methods seemed to complement each other, thus a combination of two or all three is a cost-effective solution.

Our study is limited in a number of ways. The number of test subjects in each condition was only 10 persons. This number could have been higher, but it is quite typical for method experiments. The users in the study were university students. This may introduce a bias as they may be more used to make and report assessments.

The results convey a number of interesting directions for future work. First of all, it would be interesting to try the methods out in real software projects with ordinary users. A mere replication would also be highly relevant because of the limited amount of experimental data about remote methods. This could involve improvement of each method based on our experience. The basis for comparison could also be extended with a more modern approach where the test monitor is directly involved in the identification of usability problems.

ACKNOWLEDGMENTS

The research behind this paper was partly financed by the Danish Research Councils (grant number 2106-04-0022). We are very grateful to the 40 test subjects that helped us in the usability tests. We would also like to thank the anonymous reviewers for their comments and advice.

REFERENCES

1. Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. and Stage, J. What happened to remote usability testing? An empirical study of three methods. *Proceedings of CHI 2007,* ACM Press (2007), 1405-1414.
2. Au, I., Boardman, R., Jeffries, R., Larvie, P., Pavese, A., Riegelsberger, J., Rodden, K., and Stevens, M. User experience at google: focus on the user and all else will follow. *Proceedings of CHI 2008,* ACM Press (2008), 3681-3686.
3. Bak, J. O., Nguyen, K., Risgaard, P., and Stage, J. Obstacles to usability evaluation in practice: a survey of software development organizations. *Proceedings of NordiCHI 2008*. ACM Press (2008), 23-32.
4. Brush, A. B., Ames, M., and Davis, J. A comparison of synchronous remote and local usability studies for an expert interface. *Proceedings of CHI 2004,* ACM Press (2004), 1179-1182.
5. Capra, M. G. *An Exploration of End-User Critical Incident Classification*. Master thesis, Virginia Polytechnic Institute and State University, 2001.
6. Castillo, J. C. *The User-Reported Critical Incident Method for Remote Usability Evaluation*. Master thesis, Virginia Polytechnic Institute and State University, 1997.
7. Castillo, J. C., Hartson, H. R. and Hix, D. Remote usability evaluation: Can users report their own critical incidents? *Proceedings of CHI 1998*, ACM Press (1998), 253-254.
8. Desurvire, Heather W. Faster, cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing? John Wiley & Sons, 173-202, 1994.
9. Dray, S. and Siegel, D. Remote possibilities?: International usability testing at a distance. *interactions 11*, 2 (2004), 10-17.
10. Ericsson, K. A. and Simon, H. A. How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity 5*, 3 (1998), 178-186.

11. Følstad, A., Brandtzæg, P. B. and Heim, J. Usability Analysis and Evaluation of Mobile ICT Systems. http://www.hft.org/HFT01/paper01/mobility/25_01.pdf
12. Hammontree, M., Weiler, P. and Nayak, N. Remote usability testing. *Interactions 1*, 3 (1994), 21-25.
13. Hartson, H. R. and Castillo, J. C. Remote evaluation for post-deployment usability improvement. *Proceedings of AVI 1998*, ACM Press (1998), 22-29.
14. Hartson, H. R., Castillo, J. C., Kelso, J. and Neale, W. C. Remote evaluation: The network as an extension of the usability laboratory. *Proceedings of CHI 1996*, ACM Press (1996), 228-235.
15. Hertzum, M. and Jacobsen, N. E. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction 15*, 1 (2003), 183-204.
16. Hilbert, D. M. and Redmiles, D. F. Separating the wheat from the chaff in internet-mediated user feedback expectation-driven event monitoring. *ACM SIGGROUP Bulletin 20*, 1 (1999), 35-40.
17. Kjeldskov, J., Skov, M. B. and Stage, J. Instant Data Analysis: Evaluating Usability in a Day. *Proceedings of NordiCHI 2004*, ACM Press (2004), 233-240.
18. Larsen, J. M. *Playful Interaction*. Master thesis, Aalborg University, Department of Computer Science, 2008.
19. Marsh, S. L., Dykes, J. and Attilakou, F. Evaluating a geovisualization prototype with two approaches: remote instructional vs. face-to-face exploratory. *Proceedings of Information Visualization 2006*, IEEE (2006), 310-315.
20. Millen, D. R. Remote usability evaluation: user participation in the design of a web-based email service. *ACM SIGGROUP Bulletin 20*, 1 (1999), 40-45.
21. Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J. and Stenild, S. It's worth the hassle! the added value of evaluating the usability of mobile systems in the field. *Proceedings of NordiCHI 2006*, ACM Press (2006), 272-280.
22. Nielsen, J. Finding usability problems through heuristic evaluation. *Proceedings of CHI 1992*, ACM Press (1992), 373-380.
23. Nielsen, J. and Molich, R. Heuristic evaluation of user interfaces. *Proceedings of CHI 1990*, ACM Press (1990), 249-256.
24. Nielsen, J. Usability inspection methods. *Proceedings of CHI 1994*, ACM Press (1994), 377-378.
25. Olmsted, E. and Gill, M. In-person usability study compared with self-administered web (remote-different time-place) study: does mode of study produce similar results? *Proceedings of UPA 2005*, UPA (2005).
26. Petrie, H., Hamilton, F., King, N. and Pavan, P. Remote usability evaluation with disabled people. *Proceedings of CHI 2006*, ACM Press (2006), 1133-1141.
27. Rubin, J. *Handbook of Usability Testing*. John Wiley and Sons, 1994.
28. Scholtz, J. A case study: developing a remote, rapid and automated usability testing methodology for on-line books. *Proceedings of HICSS 1999*, IEEE (1999).
29. Scholtz, J. and Downey, L. Methods for identifying usability problems with web sites. *Proceedings of IFIP Conference,* ACM Press (1998), 191-206.
30. Steves, M. P. et. al. A comparison of usage evaluation and inspection methods for assessing groupware usability. *Proceedings of CSCW 2001*, ACM Press (2001), 125-134.
31. Thompson, J. A. *Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation*. Master thesis, Virginia Polytechnic Institute and State University, 1999.
32. Tullis, T., Fleischman, S., McNulty,M., Cianchette, C. and Bergel, M. An empirical comparison of lab and remote usability testing of web sites. http://home.comcast.net/~tomtullis/publications/RemoteVsLab.pdf
33. Vermeeren, A. P. O. S., van Kesteren, I. and Bekker, M. M. Managing the 'evaluator effect' in user testing. *Proceedings of INTERACT 2003*, IOS Press (2003).
34. Waterson, S., Landay, J. A. and Matthews, T. In the lab and out in the wild: remote web usability testing for mobile devices. *Proceedings of CHI 2002*, ACM Press (2002), 796-797.
35. West, R. and Lehman, K. R. Automated Summative Usability Studies: An Empirical Evaluation. *Proceedings of CHI 2006*, ACM Press (2006), 631-639.
36. Winckler, M. A. A., Freitas, C. M. D. S. and de Lima, J. V.Remote usability testing: a case study. *Proceedings of OzCHI 1999*, CHISIG (1999).
37. Winckler, M. A. A., Freitas, C. M. D. S. and de Lima, J. V. Usability remote evaluation for www. *Proceedings of CHI 2000*, ACM Press (2000), 131-132.
38. Äijö, R. and Mantere, J. Are Non-Expert Usability Evaluations Valuable? http://www.hft.org/HFT01/paper01/acceptance/2_01.pdf