A comparison of in-lab and synchronous remote usability testing methods: Effectiveness perspective

Obead Alhadreti
Umm al-Qura University
23 PUBLICATIONS 363 CITATIONS

P.J. Mayhew
University of East Anglia
84 PUBLICATIONS 1,399 CITATIONS

Majed Alshamari
King Faisal University
34 PUBLICATIONS 788 CITATIONS

A COMPARISON OF IN-LAB AND SYNCHRONOUS REMOTE USABILITY TESTING METHODS: EFFECTIVENESS PERSPECTIVE

Mr. Obead Alhadreti[1], Dr. Pam Mayhew[1] and Dr. Majed Alshamari[2]
[1]*University of East Anglia*
[2]*King Faisal University*

ABSTRACT

Traditional in-lab usability testing has been used as the standard evaluation method for evaluating and improving the usability of software interfaces. However, in-lab testing, though effective, has its drawbacks, such as unavailability of representative end-users, high testing costs, and the difficulty of reproducing a user's everyday environment. To overcome these issues, various alternative usability evaluation methods (UEMs) have been developed over the past two decades. Among these, one of the most commonly used is the remote usability testing method. This paper is concerned with a comparative study of the traditional in-lab usability testing method and the synchronous remote usability testing method. It aims to examine how each method produces usability data, such as usability problems found, error number, time spent and success rate. The results of this paper discuss how these data differ based on the method used. It reveals some interesting results. Although the achieved data are similar, some measures differ significantly, such as identifying major usability problems.

KEYWORDS

Remote usability testing, lab testing, synchronous remote testing.

1. INTRODUCTION

Usability is increasingly recognized as an important quality factor for interactive software systems in use today. Several studies have reported the benefits of a strong commitment to usability throughout the development life-cycle of a software product. Among the observable benefits of usable user interfaces, one can mention user productivity, performance, and safety and security. Usability is important not only in increasing the speed and accuracy of the range of tasks carried out by a range of users of a system, but also in ensuring the safety and security of that use (Jean, 2004). In order to determine the usability level of a software system, a number of different but related usability evaluation methods (UEMs) have been proposed over the last three decades. One of these evaluation methods is the traditional in-lab testing method, which has been used as the standard evaluation method for evaluating and improving the usability of software user interfaces. However, traditional in-lab testing, though effective, has its drawbacks such as the availability (or otherwise) of representative end-users, the high cost of testing, and lack of a true representation of a user's environment (Hartson, Jos *et al.*, 1998). To counteract these issues, various alternatives and less expensive usability evaluation methods have been developed over the past twenty years. One such UEM is the remote usability testing method. This method addresses the above issues by relying on real users conducting a number of real scenarios in their native environments. Remote usability testing is generally classified into synchronous (moderated) and asynchronous (unmoderated) testing (Brush, Morgan *et al.*, 2004). This paper is structured as follow: it begins with exploring the current usability evaluation methods (UEMs), including recent studies related to UEM comparisons. It then discusses the paper's objective and the approach taken. Data analysis and results are also discussed. It concludes with a brief discussion and conclusion.

2. RELATED WORK AND RESEARCH OBJECTIVE

Over the last three decades, numerous research studies have been carried out in attempts to overcome usability issues in software systems. As a result, many different types of evaluation methods have been developed (see Figure 1) (Jean, 2004). Of these methods, the most frequently used are usability testing methods, usability inspection methods, and model-based methods.

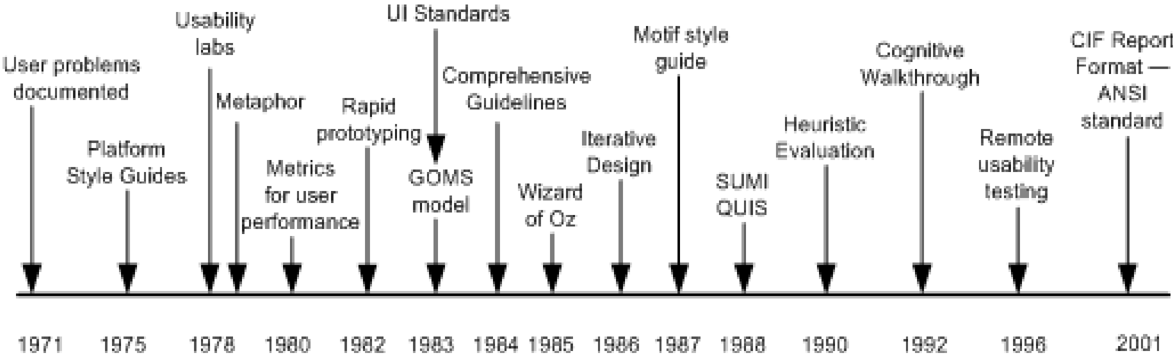


Figure 1. Development of usability evaluation methods (Jean, 2004)

Remote usability testing can be defined as, "usability evaluation where the test evaluators are separated in space and/or time from the test subjects" (Hartson, Jos *et al.*, 1996). The term "remote" refers mainly to the remote location of the test subject from the evaluator's location (Castillo, 1997). As mentioned previously, remote evaluation can be generally classified into two main categories, synchronous remote usability testing (moderated), and asynchronous remote usability testing (unmoderated) (Susan and David, 2004). Synchronous remote evaluation, sometimes referred to as 'live' or 'collaborative' remote evaluation, is a usability evaluation method that has much in common with traditional in-lab usability evaluation (Selvaraj, 2004). It involves real users participating in the evaluation process from within their own environments using their own computers. In this evaluation method, the evaluator's computer (in the usability lab) and the remote user's computer (in their natural environment) are connected in real time through the Internet using a web-conferencing or screen-sharing application, and through an audio connection via the computer or a separate phone line. These allow the evaluator to collect data on the user's actions by recording the test as the user performs the test tasks. The advantages of synchronous remote evaluation include the capability of collecting data from real users in their natural environment and the elimination of any need for participants to travel, also costs are lower making it more efficient, and more diverse users can be involved thereby including cultural contexts (Susan and David, 2004), (Morten Sieker, Henrik Villemann *et al.*, 2007). However, limited bandwidth and communication delays are some of the drawbacks of this method (Castillo, 1997).

The literature on remote synchronous testing method is more limited. Most of them conduct it by simulating in-lab settings (Susan and David, 2004), (Monty, Paul *et al.*, 1994). The rest adopted inspection methods. Tullis and others conducted research in order to evaluate the effectiveness of remote and lab usability testing methods (T. Tullis, Fleischman *et al.*, 2002). They reported the similar performance of these two methods. They found that remote testing users needed more time and completed more successful tasks than the lab group. They also received more negative feedback from those who attended lab testing (T. Tullis, Fleischman et al., 2002). However, they did not report any statistical differences between the performances of these two methods. Andrzejczak and Dahai conducted a study that attempted to clarify the effect of testing location on usability test elements such as stress levels and user experience (Chris and Dahai, 2010). Although they reported that there are no differences between the performance of remote and lab testing, their study has a number of limitations such as users' characteristics. All the users were students and the majority of them had previous experiences with the targeted website (Chris and Dahai, 2010). Another limitation is that the remote users were welcomed and briefed by the observer, which may encourage (or stress) the users to work hard in the test (Chris and Dahai, 2010). Their study shifted the focus to a new dimension in the comparisons of remote and lab testing; users' experience and stress level. The literature lacks detailed experimental methods, data analysis and empirical data, although there are a number of important attempts such as (Brush, Morgan et al., 2004), (McFadden, Hager *et al.*, 2002) and (Katherine, Evelyn *et al.*, 2004).

Although the literature usually compares two or more methods in terms of their efficiency, it should consider more whether or not the usability data are different if they are collected from different methods; remote synchronous usability testing and traditional in-lab usability testing. Such differences can mislead usability engineers who are required to produce a usability report at the end of their evaluation. Therefore, this paper examines whether or not the two methods, remote synchronous and traditional in-lab usability testing, produces different outputs. It questions the efficiency of these two methods.

3. METHOD

Two experiments were conducted in order to achieve the research objective. Two groups of users were involved. Each group consisted of 20 users to offer more validity to the comparison, as suggested by (Jacob Nielsen, 2006), (Brush, Morgan et al., 2004). This number can be also examined statistically. The users' gender, backgrounds, web experiences and ages were considered seriously in order not to influence the results later. Therefore, the characteristics of the users in each group were almost the same in terms of gender, age and Internet experience. These characteristics should reflect the targeted website audience, which is a social networking website. They all also perform the same tasks; each user performed five tasks. All the users were interested in and were familiar with the website; they participated for free. For the traditional in-lab testing section of this study, the participants carried out the test tasks in a usability lab at the university. Cam Studio screen capture software was also used to capture the participants' screen actions and record their voices on video (CamStudio, 2010). The video footages were then superimposed for further reviewing and analysis. The remote participants, on the other hand, were not provided with any equipment as they performed the test in their own environments. However, these participants were required to have a computer with Internet access, Internet Explorer 7 or higher, and a connected functional microphone. Skype Messenger 4.1 was used to connect and share the participants' screens with the researcher so that the participants' desktops could be observed for the duration of the test, and to communicate with the participants so that they could share their comments and suggestions (Skype, 2010). Although most test participants had Skype on their computers, participants who did not have Skype were advised to download it from the Internet. Before conducting a usability test, it is necessary to define clearly what metrics will be used to measure a system's usability level. According to (Sauro and Kindlund, 2005), the most frequently used usability metrics are: a) task completion rate; this concerns the percentage of tasks that are completed correctly during usability testing. Task completion rate provides a general picture of how the system being tested supports its end-users and the amount of improvement needed to make it work more effectively. b) Number of errors; this concerns the number of errors that participants make while performing the tasks. c) Time spent; this measures the time it takes a user to perform a single task from start to completion.

4. DATA ANALYSIS AND RESULTS

According to (Jacob Nielsen, 1995), usability problems can fall into one of the following categories of severity: not a usability problem, cosmetic, minor, major, and catastrophic. To ensure an objective assessment of the problems discovered in this study, the researcher sent the final set of problems identified by the two groups to a usability expert, who then classified their severity based on their frequency, impact and persistence. This classification has been used and recommended for use in usability testing in (Chen and Macredie, 2005). The following sections discuss how each of the two methods produces the usability data; number of problems discovered, task completion rate, time spent and error number. There are some similarities and differences between remote and in-lab testing. Interestingly, some of the differences were proven statistically.

Table 1 shows that both methods were able to reveal 41 usability problems. The in-lab group was slightly more effective than the remote group in identifying usability problems, as the in-lab group discovered (alone and with the remote group) almost 81% of the total number of problems, whereas the remote group only discovered (alone and with the in-lab group) almost 73% of the total number of problems. The in-lab group were able to reveal 27% of the problems discovered alone, whereas it failed to reveal 19% of the problems discovered only by the remote group. There were no statistical differences between the groups' performance except in revealing major problems. The in-lab group performed better than the remote group in revealing major problems. The Fisher exact test reveals the significance (p = .01). This means that there is 98.043% chance of the in-lab group revealing more major problems than the remote group. However, remote testing was more successful than in-lab testing in discovering minor problems. Generally, these results are in line with the findings of (Brush, Morgan et al., 2004), which concluded that in-lab testing outperformed synchronous remote testing in identifying usability problems, but they did not report any statistical differences (Brush, Morgan et al., 2004). These results are also in line with the (T. Tullis, Fleischman et al., 2002) study. They reported that lab users found more problems than remote users. In contrast, Sieker et al. reported the opposite; their results showed that synchronous remote testing out-performed in-lab testing in identifying unique major and catastrophic problems. Their findings were not sufficiently supported by statistics as the Fisher exact test did report any significant difference in the number of problems found by the two methods (p = 0.60) (Morten Sieker, Henrik Villemann et al., 2007).

Table 1. Numbers and percentages of problems discovered

The used method	Cosmetic	Minor	Major	Catastrophic	Total
In-lab uniquely	2	2	5	2	11 (27%)
Remote uniquely	3	5	0	0	8 (19%)
Both	7	6	8	1	22 (54%)
Total	12	13	13	3	41 (100%)

Each participant was asked to perform five tasks on the targeted website, meaning that a total of 100 tasks were performed by each group. At the end of each task, the researcher assessed the completion rates and then classified them as successful (completed) or unsuccessful (not completed), as suggested by (Tom Tullis and Albert, 2008). Table 2 illustrates the results derived from the measurement of the task completion rate for both groups. The participants in the in-lab group successfully completed 63 tasks out of 100, whereas the participants in the remote group were only able to complete 57. These results support the findings of a study conducted by (Morten Sieker, Henrik Villemann et al., 2007), but are in contrast to the (T. Tullis, Fleischman et al., 2002) study. A possible explanation for this difference is that the psychological effect of the physical presence of the researcher and face-to-face communication with the in-lab participants; this might have given these participants more confidence when performing their tasks. This was reported in (J Nielsen, 2005), as users tend to work harder in the lab as they feel that they are under 'test' conditions, although they were informed that the website was the target of the 'test', not themselves.

Table 2. Task completion rate for each group

Tasks	Successful	Unsuccessful	Total
In-lab	63	37	100
Remote	57	43	100

Examining these results reveals that the participants in the in-lab group worked on the test tasks more quickly than the remote group. The in-lab group spent a total of 274 minutes and 8 seconds on the test tasks, whereas the remote group spent a total of 299 minutes and 5 seconds on them. The maximum time spent on a task was 5 minutes and 47 seconds, by one of the remote participants, while the minimum time spent on a task was 45 seconds, by one of the in-lab participants. This is in line with what has been reported: users work harder in the lab as they feel they want to do it correctly within an appropriate time (J Nielsen, 2005). This difference may be due to the equipment used in each testing method. The in-lab participants all used the same equipment under the same environmental conditions, whereas the remote participants used their own computers, which led to variations in the equipment used and may have affected the time required to perform the test tasks. These results are in agreement with the findings of (Morten Sieker, Henrik Villemann et al., 2007), (T. Tullis, Fleischman et al., 2002)and (Chris and Dahai, 2010), which show that in-lab testing participants complete tasks more quickly than remote testing participants. Figure 2 below compares the average time spent by each group on each individual task.

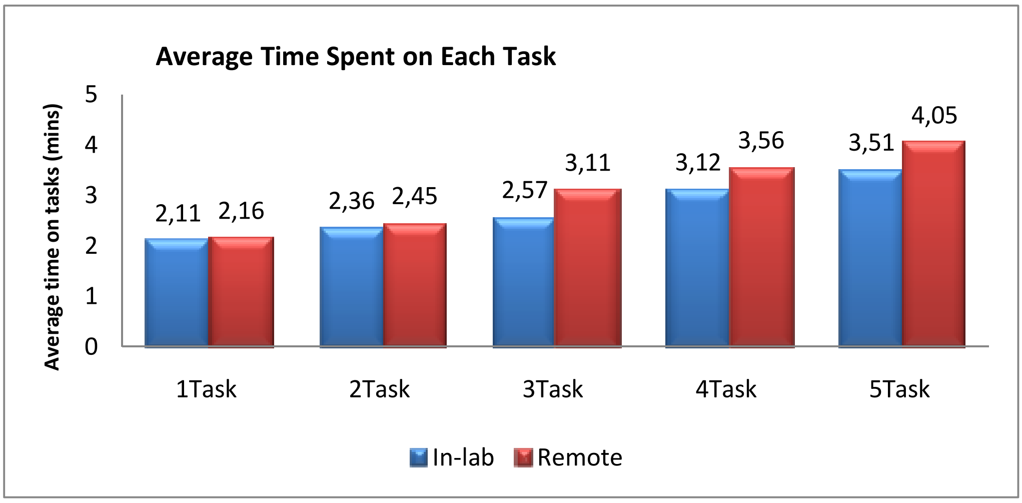


Figure 2. Average time spent on tasks

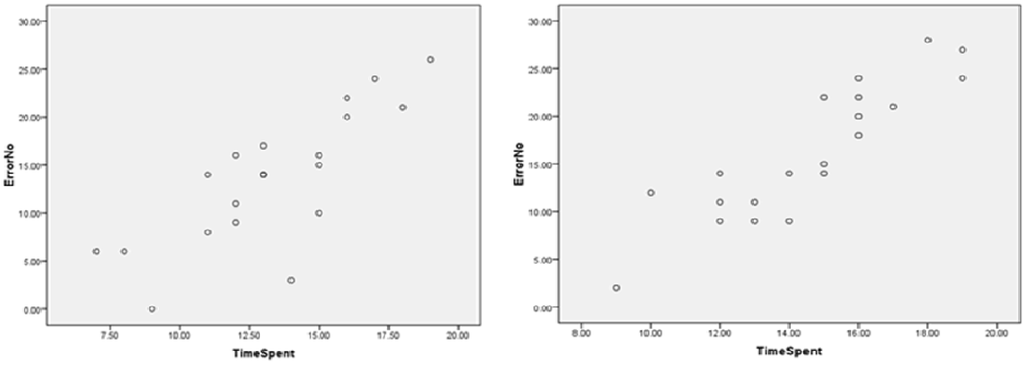
Table 3 shows the number of errors participants made on each task, the total number of errors on all tasks, and the average number of errors made by each participant.

Table 3. Number of errors on each task

UEMs	Task 1	Task 2	Task 3	Task 4	Task 5	Total	Average
In-lab	21	32	59	73	87	272	13.6
Remote	28	46	62	88	102	326	16.3

The total number of errors recorded for the remote group is larger than that recorded for the in-lab group. The difference between the in-lab and the remote groups' recorded errors might be due to the fact that, in the lab workstation, the researcher could ensure that the health, safety and ergonomic requirements for computer use were applied, whereas the researcher was unable to do the same for the remote group. Another reason might be that the in-lab participants were more concerned with how they would be judged by the researcher, who was located with them in the same place, hence they have tried to concentrate harder on the tasks. In general, these results support the findings of (Katherine, Evelyn et al., 2004), which concluded that remote testing participants make more errors than in-lab testing participants whilst performing tasks. The results in this section all suggest that the participants in the in-lab group were slightly more successful, efficient and accurate than the participants in the remote group, with regard to completing the test tasks.

In this research, it has been found that the users who spent more time, made more errors (in both groups), which can be seen clearly in Figure 3 below. This has been examined statistically and it was found that there is a strong statistical relationship between the time spent by users and the errors made by them (p = .001). The p value is the same for each group and for the two groups together.



Lab testing

Remote synchronous Testing

Figure 3. The strong correlation between time spent and errors made by users

5. DISCUSSION

Reliable study needs solid and clear steps. This study follows a systematic approach; one that is recommended for use in usability studies. All the usability factors (users number, characteristics, tasks number, the targeted website and usability measures) were taken into consideration in order to eliminate any influences that may occur or affect the achieved results. There are some interesting results; this study found significant differences between in-lab and remote usability testing in discovering major usability problems. No other differences were found between the two methods in terms of discovering catastrophic, minor and cosmetic usability problems. The results show that the two groups (i.e. methods) performed similarly, except that the in-lab group performed better than the remote group in discovering major problems; this was proved statistically. These results are in line with (Brush, Morgan et al., 2004), although they did not prove theirs statistically. The achieved results conflict with the (Morten Sieker, Henrik Villemann et al., 2007) study. The reasons behind this difference need further examination, such as investigating the most influential usability factors (test environment, tasks, user number, observer, the targeted website, users' characteristics and others).

The remote usability testing group needed more time, made more errors and performed fewer successful tasks than the in-lab group, and these results are in line with (Chris and Dahai, 2010), (Morten Sieker, Henrik Villemann et al., 2007) and (Katherine, Evelyn et al., 2004). There may be various reasons for the remote group's poorer performance, possibly related to those users' equipment such as their machines and Internet speeds. The other possible interpretation is that users tend to work harder in the lab than in normal circumstances (J Nielsen, 2005). If this is the case, the achieved results of the in-lab testing group may mislead usability engineers over the usability of the website. However, these performance differences (time, errors and success) were not proven statistically; further experiments are needed to clarify the reasons for these differences and to justify them.

6. CONCLUSIONS

In our comparison, we found some differences between in-lab and remote synchronous usability testing. We saw that there is a statistical difference between these two groups in terms of identifying major usability problems. The in-lab group revealed more major usability problems than the remote synchronous group. However, the in-lab group spent less time, made fewer errors and performed more successful tasks than the remote synchronous group. No statistical differences were reported for these differences. This research suggests that although the levels of efficiency for both methods are almost the same, other aspects should be investigated, including test cost, time and ease of application. Further investigation of these aspects, including method effectiveness, would be to the benefit of the website design sector.

REFERENCES

Brush, *et al.* (2004). A comparison of synchronous remote and local usability studies for an expert interface. CHI '04 extended abstracts on Human factors in computing systems. Vienna, Austria, ACM.

Camstudio (2010). CamStudio. http://camstudio.en.softonic.com**:** Screen capture software.

Castillo (1997). The User-Reported Critical Incident Method for Remote Usability Evaluation, Virginia Polytechnic Institute and State University. MSc.

Chen and Macredie (2005). "The assessment of usability of electronic shopping: A heuristic evaluation." Internationa journal of Information Management 25: 516-532.

Chris and Dahai (2010). "The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience." J. Syst. Softw. 83(7): 1258-1266.

Hartson, *et al.* (1998). Remote evaluation for post-deployment usability improvement. Proceedings of the working conference on Advanced visual interfaces. L'Aquila, Italy, ACM.

Hartson, *et al.* (1996). Remote evaluation: the network as an extension of the usability laboratory. Proceedings of the SIGCHI conference on Human factors in computing systems: common ground. Vancouver, British Columbia, Canada, ACM.

Jean (2004). "Usability evaluation." Encyclopedia of Human-Computer Interaction IAD National Institute of Standards and Technology

Katherine, *et al*. (2004). Here, there, anywhere: remote usability testing that works. Proceedings of the 5th conference on Information technology education. Salt Lake City, UT, USA, ACM.

Mcfadden, *et al.* (2002). "Remote Usability Evaluation: Overview and Case Studies." International Journal of Human-Computer Interaction 14(3): 489 - 502.

Monty, *et al.* (1994). "Remote usability testing." interactions 1(3): 21-25.

Morten Sieker, *et al.* (2007). What happened to remote usability testing?: an empirical study of three methods. Proceedings of the SIGCHI conference on Human factors in computing systems. San Jose, California, USA, ACM.

Nielsen (1995). Severity Ratings for Usability Problems. www.useit.com.

Nielsen (2005). Authentic Behavior in User Testing. www.useit.com.

Nielsen (2006). Quantitative Studies: How Many Users to Test? www.useit.com.

Sauro and Kindlund (2005). A Method to Standardize Usability Metrics Into a Single Score. Conference on Human Factors in Computing Systems, Portland, Oregon, USA.

Selvaraj (2004). Comparative Study of Synchronous Remote and Traditional In-Lab Usability

Evaluation Methods, Virginia Polytechnic Institute and State University. MSc.

Skype (2010). Skype. http://www.skype.com/intl/en-gb/get-skype/on-your-computer/windows**:** Screen capture software.

Susan and David (2004). "Remote possibilities?: international usability testing at a distance." interactions 11(2): 10-17.

Tullis and Albert (2008). Measuring the user experience. Burlington MA, Elsevier Inc.

Tullis, *et al.* (2002). An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. Usability Professionals Association Conference.