Users' Performance in Lab and Non-Lab Enviornments through Online Usability Testing

A Case of Evaluating the Usability of Digital Academic Libraries' Websites

Abeer Alharbi and Pam Mayhew
University of East Anglia
Norwich, UK
Abeer.Alharbi@uea.ac.uk, P.Mayhew@uea.ac.uk

***Abstract*—The factors related to the environment in which users operate may be of a vital importance when trying to understand how they experience a particular system. It is required that we find out how we can get to know those factors to investigate if they affect the users' performance in usability testing. An online usability study has emerged that can be attempted by a large, varied pool of users' anywhere with an Internet connection. Would the usage of an online usability study help to give comprehensive insight and an understanding of the whole user experience? That is especially interesting if the user operates remotely, as we are unaware of what the users might experience while performing the test (e.g., distractions and type of device used to attempt the test). Accordingly, a pilot study of ongoing research was conducted. An identical online usability-testing tool (Loop11) through which to apply the online usability study was used in two environments: unrestricted (the user's natural environment), and totally restricted (a simulated lab environment). Ten subjects completed the test in the restricted environment and 20 completed it in the unrestricted environment. All of the subjects were asked to perform predefined search tasks on digital libraries' websites. Their performance was analyzed and compared against the two different environments. The results showed that online usability testing is a feasible method to gain comprehensive insight into how users attempt usability testing in a non-lab environment. The results of whether different environments affect test performance show no valuable differences in most of the study's measurements. The test subjects were frequently multitask while they performed the usability testing in an unrestricted environment, but they were highly distracted if they personally interrupted. The results encourage the researcher to conduct a formal version of this study to further examine the learned lessons from the pilot study.**

Keywords—environment; distraction; usability testing; online; remote

I. BACKGROUND

Usability testing is normally carried out in controlled usability lab environment to assess the system usability. It is traditionally based on evaluator observations of the user [1]; all interactions are recorded and observed [2].

A traditional lab test involves a small number of users with a high commitment to perform the tests – conditions that sufficiently enable the detection of the most obvious usability issues [1]. As a result of the advances in networking and communications technologies, synchronous and asynchronous methods have supported the remote application of the usability evaluation techniques.

The benefits of a remote usability evaluation are driven by the following motivations:

- This method has now being used extensively because it allows for the recruitment of many users, thereby reducing travel time and cost.
- It allows for testing in the user's natural environment [3]. A large pool of participants can comfortably take part in a test administered in a familiar setting [4].

We will focus on the empirical work of assessing the usability testing that is done asynchronously; it is similar to the online usability testing in that the user and evaluator are separated by both time and space. Some researchers concluded that asynchronous methods are more time-consuming for participants and enable the identification of fewer usability problems [6], whereas others affirm the appeal of such methods as alternatives to traditional lab usability testing. Although fewer usability problems are identified, there is significantly less time involved in such approaches [8].

Under asynchronous methods, a higher number of diverse problems are reported by participants than those determined in traditional lab usability testing [9]. With remote asynchronous method-based studies, the participant recruitment is also simpler; a larger number of participants and a more diverse sample can be recruited because they will not be required to spend any time away from their regular activities.

A few researchers described how they asynchronously collected data. Most remote studies centered on the spatial and temporal differences between the evaluators' and participants' environment, focusing on pros and cons of lab versus remote usability testing, together with practical recommendations for an improved remote usability method reporting from the empirical application of that method versus the traditional usability evaluation method(s).

Most of the studies have neglected the environmental factors that these new forms of the usability testing (remote testing) can impose on the generated results.

Discussing the characteristics of the different usability testing methods indicates the need for methods that provide as much insight as possible into the usability considerations of

both the user's and system's perspectives. Such methods should be time- and cost-efficient.

Online usability testing has emerged and can be attempted anywhere by any Internet connection; it employs a complete user experience by allowing the involvement of a large and varied pool of participants [9].

The term 'online' describes the test environment. When this term is combined with 'remote', it describes the form of connection between the evaluator and the participant in this environment [10]. Online studies do not impose new methods; they adopt traditional methods with a new view of the research design [10]. Shall online usability studies be adopted to meet that need? Such possibility points to a rich field of investigation.

II. Related Work

Ref [7] examined the effect of the test location on the usability testing performance, participant stress level, and subjective testing experience. The only significant difference between the synchronous testing and the lab-based usability testing is the task time; there was no significant reported difference in the stress level and number of critical incidents or subjective assessment found. However, Ref [7] did not address in its study the environmental factors, other than the location, that may have affected the performance.

Ref [11] used children as the respondents when examining the effects of the testing environment on the results of a usability evaluation process using field testing. The authors concluded that the field testing could be a viable approach to reducing the time used to complete a given task, and to minimising the frustration levels reported by children during such tests. This study used filed testing, which is restricted to certain usability testing settings with certain user samples.

Ref [10] used an online usability study to investigate the existence of distraction during online user studies in digital libraries and analysed their influence. The same test was set up using Loop11 in the lab and in the user's natural environment; these tests were completed asynchronously by different groups of participants. The results showed that the participants who were in their natural environment were highly distracted and needed more time to complete the test. The test environment did not affect the successful task completion, the participants' judgments of the websites, or their decision-making processes. Multitasking, which seemed to be the obvious influencing distraction in the natural environment, did not increase the time score in a significant way. However, Ref [10]'s study did not allow users to report or rate usability problems. The participants gave only ratings for the websites' usability and task complexity. The factors related to the users' environment for usability testing need more invistigation.

III. Pilot Study Aim and Objectcives

This exploratory research has multiple aims, and in this pilot study we investigates the feasibility of an online usability study though the usability testing in users' natural environment to collect data related to that environment. To that point, this pilot study investigates the effectiveness of online usability studies in revealing comprehensive insight into the evaluation of the website's usability, with consideration given to the different factors related to user environment. The presented work has the following objectives:

A. To investigate the effectiveness of online usability studies in providing data on the test performance

B. To investigate the differences in user performance between online testing in labs versus natural environments

Detailed theorizing and empirical investigation could be followed in [5], which are mainly summarized by the findings of [6] [7] [8], which indicate that usability testing performance in a lab environment was better than that in a user's natural environment. It was expected that the participants' test performance in the two environments would be better in the lab environment. However, building on the findings of [11], it was presupposed that the online usability study would enable the collection of data on the following measures: 1) time, 2) self-reports, 2) click stream, 3) questionnaires, 4) effectiveness in terms of successful task completion, 5) overall success 6) and visited URLs. The researcher had no previous knowledge of whether an online usability study enabled the users to self-report the data.

IV. Methodolgy

1) Test Environment, Setting, and Procedures

The experiment lasted 14 days and was conducted in two different environments, as shown in Fig. 1. The first environment was unrestricted (conducted in each subject's ordinary natural environment at a time of his or her choice) using any device or communication technology with the user most likely exposed to distractions. The second was a restricted, distraction-free lab environment, and used a network and apparatus controlled by the University of East Anglia (UEA). The usability testing session was conducted in a quiet room in the university library to simulate artificial labs.

In the first four days, the participants were recruited for the unrestricted usability testing with e-mails, Facebook, Twitter, and advertisements on the university's bulletin boards. The e-mail messages introduced the study and contained a link to the study portal. The study itself was re-introduced in the usability website portal (a self-designed web portal that contained introductory information to the test, instructions, and contact information). The designed web portal was useful in case the subjects forgot any information or needed additional information (e.g., Twitter does not allow posts of more than 140 characters, so only the invitation and link were provided).

A direct link to the test was not provided in the initial message to prevent users from choosing the participation site in advance to the test. However, this approach did not generate a good response. Recruitment continued for ten more days in the same way, but a direct link was included.

The subjects within the restricted environment were recruited using distribution recruitment flyers placed throughout the UEA campus; the flyers indicated the library room number and test location and contained the same information as the emails and Facebook page but without the study URL. In the room, the portal URL was shown on a sheet of paper next to the PC where the test was administered. The

test's URL was not included on the recruitment flyer to avoid the instance of potential users taking the test in an uncontrolled environment (e.g., home). Only one browser was installed on the UEA PC on which all the restricted environment samples were recorded.

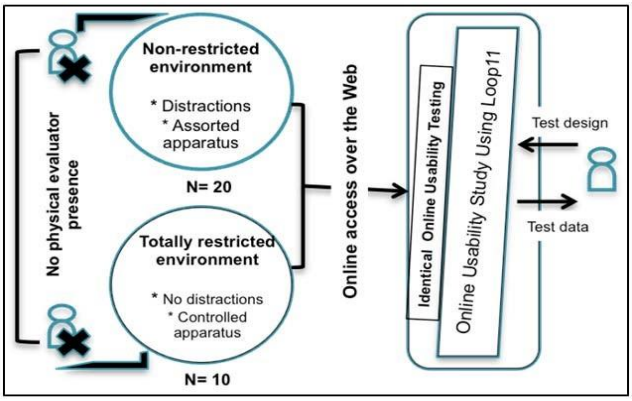

Fig. 1. Pilot testing's Design Specifications

We claim that bias was avoided in this setting because the subjects were not told if there was other setting for the usability testing. In addition, the participants were not told that the main purpose of the test was to evaluate the usability of the digital library or to evaluate their performance; rather they were asked to perform the tasks as they would normally perform them. This should reduce the possibility of users being affected by the study's purpose and thus having their performance affected.

The subjects in both environments were unaware they were being timed. The questions pertaining to the distractions and settings were placed at the end of the test. The subjects in the unrestricted environment were told that they could take part in the test when it was convenient for them.

No guidelines were provided to the subjects regarding the idea of multitasking or interruptions in the unrestricted environment. However, the restricted environment subjects were asked to not be distracted while carrying out the test; this was mentioned in the recruitment flyers.

The web portal was accessed by the test subjects in their respective environments; consent by the subject was required given prior to he/she being directed to the Loop11 usability testing website, which opened in a new window. If the participants had questions, they could visit the FAQ page in the designed web portal. They were also advised to contact the researcher via study's email in case they faced any problems or needed assistance.

2) Test objects:

Usability testing studies usually cover a single search engine or a few predetermined test objects. However, this study was not intended to evaluate a specific website, but it aimed to investigate the effects of different environments on usability testing performance. Three digital libraries were used as test objects. Amazon.co.uk served as a control website, and JSTOR was incorporated to train the users on testing tasks at the beginning of the test. The other websites were CiteSeerX, Perseus and arXiv, as shown in Table 1. These digital libraries were selected after their specialties and interfaces' design had been investigated. The digital libraries were tested by the researcher and were found to have various usability issues.

TABLE I. TEST'S OBJECTS AND TASK DIFFICULTY LEVEL

	Test Object	**Difficulty Level**
Training Task	JSTOR	Simple
1st task	CiteSeerX	Simple
2nd task	Perseus	Difficult
3rd task	arXiv	Difficult
4th task	Amazon (Control website)	Simple

3) The Test Study Design, Tasks, and Questions:

Participants were asked to search for a specific document (e.g., file) or information on the websites. All of the digital libraries were fully functional during the test window. The participants' confidence with the kinds of digital libraries and the types of tasks was high because the tasks resembled their preparations for essays or class papers. For the analysis, it was important to use at least one well-known website to determine whether previous knowledge of the site changed the behavior. This control website needed to be similar to the previous digital libraries in that it allowed a similar search task. Amazon was chosen as a control website because it allowed the users to search for a book, had permanent URLs, and provided search results that were relatively stable compared to eBay or Google, for example [10].

Fig. 1 shows that the two groups that utilized the same usability testing tool (Loop11) whether the test location was in the lab or in the participant's natural environment exposed to distractions and/or allowed to use only the standard apparatus.

The usability testing Web site presented a task-based interface that the user navigated according to his/her choices among options (e.g. 'task complete', 'task abandon', or 'continue'). Initially, each participant was asked to perform a non-timed training task using the JSTOR digital library website. The training exercises were designed to familiarize the users with Loop11 and the nature of the test tasks. The subjects were told that they did not need to provide answers to the training tasks. Then, they were asked to perform the actual timed tasks that included a search in the digital libraries, including CiteSeerX, Perseus, arXiv, and Amazon.

The subjects were instructed to choose the 'task complete' option once they believed they had retrieved the required information. However, if the information retrieved was incorrect, the task was considered a failure. If the subject recognized that he/she was unable to find any appropriate information, 'abandon the task' should be selected, Fig. 2 shows the variables defined to collect the data for each measurement.

In this experiment, each task required obvious and assessable endpoints. However, the practicality of the subjects' self-reporting task completion was restricted, as Loop11 requires accurate URLs to outline the success or failure of any search. Loop11 can only provide information based on working URLs. Therefore, if a user does not find the correct URL predefined by the study designer, Loop11 will consider the

corresponding task a failure. As a result, participants should be instructed to look for tangible information from the document, and hints have been provided for all the tasks so they can recognize whether they should choose the 'task complete' or 'abandon the task' option (self-administered checks [8]). This allows the user to abandon the task without embarrassment when he or she is unable to provide further information on the task.

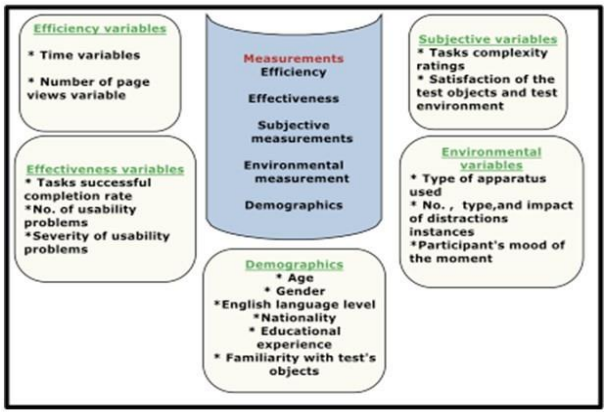

Fig. 2. Usability Testing Performance's Measurements and Their Contributed Variables

Some tasks were more difficult than others because of the difficulty of approaching the target information, and sometimes the participants needed advanced skills to deal with poorly designed Web sites. In addition, some users lacked the necessary background to make sense of some of the tasks. To address this problem, all of the participants worked on tasks of various levels of complexity.

After completing each task, the test subjects were asked to rate task complexity and to assess website functionality and usability using 5-point Likert scale based questions. Then, they were asked to report if they had encountered any usability problems while performing the test and, if so, to describe it/them and assess their severity (critical, serious, cosmetic). After completing the tasks, they were asked about their familiarity with the websites used in the test and whether they encountered any technical problems. They were asked to report whether they were multitasking during the test or were interrupted. Questions about their internal distraction (mood of the moment, level of interest, not-related thoughts) were also asked. If a test subject experienced any type of distraction, he/she was asked to assess the effect of that distraction on his/her performance. The participants were urged to provide honest answers to ensure that accurate data collection and were assured that their answers would not affect their participation reward; the purpose of this disclosure was to avoid social desirability responses. Questions about the test subjects' demographics were asked at the end of the test.

4) Test subjects:

A between-subjects design was used and the experimental usability testing included 30 different subjects recruited from various UEA facilities; 20 of the subjects served as the unrestricted test environment sample; the other 10 served as the restricted test environment sample. Their education ranged from the undergraduate to the PhD levels. The experiment was completed in totally unrestricted and totally restricted environments.

V. Pilot Study Analysis and Results

A. The Effectiveness of Online Usability Studies in Providing Data on Test Performance

1) Reporting Usability Problems

As indicated in the methodology section, this type of data was claimed from the test subjects. If the test subject indicated that he/she encountered usability problem(s) in a task, he/she was asked to report it/them. All of the test subjects who indicated usability problem(s) reported/described them.

Ninety-nine percent of the test's subjects in the unrestricted environment indicated and described that they encountered usability problems; 90% of the test's subjects in the totally restricted environment indicated this, as well, as shown in Table 2. The Fisher exact test showed no significant association between the type of test environment and whether the participants reported usability problems in the entire test (P=1.0).

TABLE II. Reporting Usability Problems

	Reporting usability problems	
	Unrestricted environment	***Totally restricted environment***
Percentage (%) of the test subjects' in certain environmnents who reported/desrcibed usability problems in the test	99%	90%
Fisher's exact test	No significant association between the type of test environment and whether participants reported usability problems in the entire test (P=1.0)	

2) Reporting Severity of Usability Problem

Of all the test subjects who reported that they encountered usability problems during the test session, 83.8% reported ratings for those problems' severities.

Of the problems identified in the unrestricted environment samples, 13.1% of the severity ratings were unreported; 3.8% of the severity ratings in the restricted environment samples were unreported. The result of Fisher's exact test indicated that there was no significant association between the test's environments and whether the participants reported the severity of the usability problems they identified in the test (p = 0.378), as shown in Table 3.

TABLE III. Reporting Severity Ratings of Usability Problems

	Reporting Severity Ratings of Usability Problems	
	Unrestricted environment	***Totally restricted environment***
Percentage (%) of the test subjects' who did not report the severity rating of the usability problems across the environments	13.1%	3.8%
Fisher's exact test	There was a significant association between the type of test environment and whether the participants reported the severity rating of the usability problems in the entire test (P=0.378)	

3) Reporting Distraction Instances (Unrestricted environment samples only)

Distraction instances is a type of data that was claimed only from the test subjects in the unrestricted environment. All of the test subjects who indicated that they experienced distraction instances during the test reported them; 64.3% of them were distracted by multitasking (working with the same machine but with task[s] other than the test) and 45% were distracted by personal interruptions (being contacted by a phone call, SMSs, or other distractions not related to the same machine), as shown in Table 4.

TABLE IV. REPORTING DISTRACTION INSTANCES IN THE UNRESTRICTED ENVIRONMENT

	Reporting Distraction Instances	
	Multitasking	***Interruption***
Percentage (%) of the test subjects, in unrestricted environments who were distracted by this type	13.1%	3.8%
Average and standard devation of the distractions caused by each type, sample size (AVG:STD, N)	(1.78:1.1, 9)	(2.13:1.13, 8)

4) Reporting Type of Used Apparatus (Unrestricted environment samples only)

All the test's subjects in the unrestricted environment reported types of apparatus that have been used during the test.

B. To Investigate the Differences in User Performances between Online Testing in Labs and in Users' Natural Environments.

1) Efficiency:

a) Time Scores

Table 5 shows that the mean (average) value for the time score in the unrestrcited environment is larger than those in the restricited environment. The Mann-Whitney U test showed that there is no significant difference in the time required per task (taskt, task2, task3, and task4) between the test subjects in the test environments. This was also true when the values of each task were added together (time for all tasks). However, there was a significant difference between the two test environments. The time for the whole test was affected because it is the sum of the time for all of the tasks and the time per question all together.

b) Page Views

Another measurement of efficiency in this study is how many webpages were viewed in the test session to complete the test. The webpages were limited to those of the test object (the websites of the digital academic libraries disserted to this test). These data are captured and recoreded by Loop11. We notice that most of statistics listed in Table 6 indicate larger mean values of page views in the restricted environment. Yet, the Mann-Whitney U test showed that the number of pages viewed by participants in all of the restricted environment test tasks (Mdn = 23) did not differ significantly from the values in the unrestricted environment (Mdn = 19.50), U = 61.500, z = −1.697, p = 0.91 and r = −0.3. However, the effect size r was considered small.

TABLE V. TIME SCORES STATISTICS

	Task Completion Time (in seconds) (Average: STD)		**Mann-Whitney U test**
	Unrestricted environment	***Totally restricted environment***	
Task1	(120.50:53.44)	(89.70:30.76)	(U=126, p=0.164, r=0.3)
Task2	(107.73:48.18)	(96.40:47.53)	(U=126, p=0.164, r=0.26)
Task3	(232.10:146.1)	(222.90:123.87)	(U=113, P=0.588, r=0.11)
Task4	(147:59.1)	(90.67:21.24)	(U=141, p=0.005, r=0.52)*
Per all tasks	(620.11:245.51)	(507.60:140.21)	(U=119, p=0.175, r=0.26)
Questions	(1161.23:335.95)	(562.71:311.82)	(U=149, p=0.000, r=0.7)*
Entire test session	(1572.559:424.6)	(1099.67:154.06)	(U=131, p=0.002, r=0.6)*

*Indicate significant difference

TABLE VI. PAGE VIEWS STATISTICS

	Page views (Average: STD)		**Mann-Whitney U test**
	Unrestricted environment	***Totally restricted environment***	
Task1	(3.83:1.724)	(3.50:0.926)	(U=74, p=0.935, r=0.2)
Task2	(4.17:0.514)	(4.90:1.370)	(U=63, p=0.248, r=-0.3)
Task3	(5.40:3.548)	(7.70:3.974)	(U=62.500, p=0.138, r=-0.3)
Task4	(4.50:1.762)	(5.20:1.989)	(U=86.500, p=0.559, r=-0.1)
Per all tasks	(19.15:6.072)	(23.10:5.859)	(U=61.500, p=1.697, r=-0.3)

2) Effectiveness

a) Successful Completion per Task

There was no significance association between the type of test environment and whether task1 completed successfully using Fisher's exact test (p=0.235>0.05). This is also true for task2 (p=1.000>0.05), task3 (p=0.251>0.05), and task4 (p=0.640>0.05), as shown in Table 7.

However, the Mann-Whitney U test for significance showed that the number of successfully completed tasks in the

restricted environment (Mdn = 3) did not differ significantly from that in the unrestricted environment (Mdn = 2), U = 59, z = −1.95, p =0.74 and r = −0.4. Yet, the effect size r was considered small.

TABLE VII. SUCCCESSFUL COMPLETION OF TASKS

	Percentage (%) of test's subjects who successfully completed the task within test environment (Average: STD)		**Fisher's exact test**
	Unrestricted environment	***Totally restricted environment***	
Task1	50%	80%	p=0.235
Task2	85%	90%	p=1.000
Task3	5%	20%	p=0.251
Task4	80%	90%	p=0.640

b) Number of Usability Problems

Ninety-nine problems have been identified in the test. Thirty of these problems were identified in the restricted environment (33.33% of the all problems reported in the test) while 69 problems were reported in the unrestricted environment (69.69% of the all problems reported in the test).

The Mann-Whitney U test showed that the number of problems identified in the restricted environment session (Mdn = 3) did not differ significantly from that of the unrestricted environment session (Mdn = 3), U = 107.5, z = 0.338, p = 0.746 and r = 0.7. However, the effect size r was considered medium, as shown in Table 8.

TABLE VIII. STATISTICS OF USABILITY PROBLEMS

	(AVG: STD) of usability problems per task, (MIN, MAX) per participant, sum per task		**Mann-Whitney U test**
	Unrestricted environment	***Totally restricted environment***	
Task1	(0.6:0.8), (0, 2), 12	(0.8:1.14), (0, 3), 8	(U=94.00, P=0.812, r=-0.1)
Task2	(0.40:0.7), (0, 2), 8	(0.3:0.9), (0, 3), 2	(U=117.00, P=0.475, r=0.2)
Task3	(1.70:1.4), (0, 5), 43	(1.6:1.08), (0,3), 18	(U=102.00, P=0.984, r=0.2)
Task4	(0.30:0.5), (0, 1), 6	(0.3:0.5), (0, 1), 2	(U=100.00, P=1.000, r=0.0)
Per all tasks	**(AVG: STD) per test, (MIN, MAX) per participant, Sum per test**		**Mann-Whitney U test**
	(3.5:1.8), (0, 7), 69	(3:2.2), (0, 8), 30	(U=107.500, P=0.746, r=0.7)

c) Agreed Usability Problems and Their Severity Ratings within the Same Test's Environment

The test subjects in the unrestricted environment tended to identify a greater number of problems that belong to the same category according to categories [1] of usability problem specified by [12] and of prcoblems that agreed in their conent (indicated similar thing). The restricted environment test's subjects tended to give more critical ratings (rating scores that approached a critical value of 3) than those in the other environment. However, this might have been a questionable determination since a large number of the problem ratings in the unrestricted environment were unreported, as shown in Table 9.

TABLE IX. AGREED USABILITY PROBLEMS AND THEIR SEVERITY RATINGS WITHIN THE SAME TEST'SENVIRONMENT

	Number of problems agreed on the belonging-to category	**Number of problems agreed on their content**	**Severity rating of problem agreed on the belonging-to category (AVG: STD)**	**Severity rating of problems agreed on the their content (AVG: STD)**	**Number of unreported ratings**
	Restricted Environment				
Task1	8	5	(2.6:0.8)	(2.7:0.4)	0
Task2	2	1	(1:0)	NA*	0
Task3	9	11	(3.0:0.0)	(3:0)	2
Total	19	17	(2.8:0.9)	(2.7:0.9)	2
	Unrestricted Environment				
Task1	9	5	(2.0:1.2)	NA*	6
Task2	8	5	(1.3:0.4)	(1.4:0.5)	3
Task3	43	41	(2.4:0.6)	(2.4:0.7)	4
Total	60	51	(2.2:0.9)	(2.3:0.9)	13

**NA indicates insufficient/missed numerical sources to apply such statistical procedure on.*

d) Agreed on Usability Problems and Their Severity Ratings between the Two Test Environments

The unrestricted test subjects identified a greater number of problems in all of the agreement instances between the identified usability problems. Yet, this group tended to underreport the severity ratings of these problems; 19.1% of the severity ratings of the agreed-on problems[2] identified in the unrestricted and restricted environments were not reported compared to 9.1% missed ratings by the restricted participants; for the details of the related statistics, refer to Table 10.

a) Unique Usability Problems

The percentage of unique problems of those identified in the unrestricted environment was greater than the percentage of the corresponding ones in the restricted environment (8.7 > 3.3%), as shown in Table 11. The percentage is used here because the sample size of the test subjects in the two environments was not equal.

[1] After classifying the usability problems reported by test's subjects. Classified problems were then assigned to the usability problems categories defined by [12].

[2] For briefness, only if the problems agreed on their content were they considered and counted within an agreement instance.

TABLE X. AGREED USABILITY PROBLEMS AND THEIR SEVERITY RATINGS BETWEEN THE TWO TEST ENVIRONMENTS

	ATA[1] (No.)	AP[2] Total	*Unrestricted environment*			*Totally restricted environment*		
			PPA[3] (AVG)	SRPPA[4] (AVG: STD)	UP[5] (No.)	PPA[3] (AVG)	SRPA[4] (AVG: STD)	UP[5] (No.)
Task1	2	5	1.5	NA	1	1	NA	0
Task3	3	18	3.6	2.3	3	2.75	(3:0)	1
Total	5	23	2.8	2.3	4	2.1	(3:0)	1

1. Number of any-agreement instances
2. Total number of agreed problems
3. Average number of problems per agreement
4. Average: standard deviation of severity rating of problems per agreement
5. Number of reported ratings

TABLE XI. UNIQUE USABILITY PROBLEMS

	Unrestricted environment	*Totally restricted environment*
Task1	3	1
Task2	3	1
Task3	1	0
Total	**(No[1], %[2])**	
	6 (8.7%)	1 (3.3%)

1. Total number of unique usability problems in each type of environment.
2. Percentage of unique problems with respect to the total usability problems identified in each test environment.

b) Frequency of a Specific Severity Rating

The Mann-Whitney U test showed that there was no significant difference between the number of usability problems when classified in terms of their severity ratings between the two test environments; refer to Table 12 for more statistics.

3) Subjective Ratings

a) Task Difficulty

The Mann-Whitney U test showed that there was no significant difference in task complexity ratings between the two environments; refer to Table13 for statistics.

b) Test Objects' Usability

The Mann-Whitney U test showed no significant difference among the subjective ratings given for all aspects[3] of the Web sites between the two groups.

c) Satisfaction about the Test Location and Setting

There was no significant difference between optimum scores given for satisfaction in test location and setting (Mdn = 2) in the restricted and unrestricted environments: U = 62, z = −1.418, p = 0.199 and r = −0.3. The effect size r was considered small.

TABLE XII. SPECIFIC SEVERITY RATING STATISTICS

	(AVG: STD) per task, sum per task					
	Unrestricted environment			*Totally restricted environment*		
	Critical	**Serious**	**Cosmetic**	**Critical**	**Serious**	**Cosmetic**
Task 1	(0.15:0.366), 3	NA	(0.10:0.308),2	(0.60 : 0.00), 6	(0.10:0.32), 1	NA
Task 2	NA	(0.05, 0.224), 1	(0.20, 0.523), 4	(0.20:0.632), 2	NA	(0.10: 0.316), 1
Task 3	(0.85:0.745), 17	(0.95: 1.317), 19	(0.15, 0.366), 3	(1.30: 1.509), 13	(0.10:0.316),1	NA
Task 4	(0.10:0.0.308), 2	(0.05: 0.224), 1	(0.15: 0.366), 3	NA	(0.10:0.316), 1	(0.10: 0.316), 1
All tasks	(1.10:1.119),22	(1.05: 1.317), 21	(0.60:0.995), 12	(2.10:1.449), 21	(0.30:0.483), 3	(0.20: 0.422),2

NA indicates here that no such severity rating was given or this task in this environment.

TABLE XIII. STATISTICS OF TASK DIFFICULTY RATINGS

	(Average: STD)		**Mann-Whitney U test**
	Unrestricted environment	**Totally restricted environment**	
Task1	(1.50:0.889)	(1.4:0.699)	(U=102, p=0.948, r=-3)
Task2	(1.85:1.226)	(2.20:1.619)	(U=95, p=0.846, r=-3)
Task3	(3.95:1.508)	(3.90:1.792)	(U=85, p=0.668, r=-0.3)
Task4 (with control test object)	(1.26:0.452)	(2.200:0.47)	(U=98.500, p=0.875, r=-0.3)

[3] Aspects include the search's results relevancy, search function goodness, Website overall usability, Website design, and Website overall goodness.

TABLE XIV. SATISFACTION OF RATINGS GIVEN FOR TEST LOCATION AND SETTING (RATING SCORES: [1] VERY GOOD – [5] VER BAD)

	(Average: STD)		Mann-Whitney U test
	Unrestricted environment	*Totally restricted environment*	
Satisfaction about test location and setting	(1.80:0.77)	(2.22:0.67)	(U=62, p=0.199, r=-0.3)

4) Environmental Factors

a) External Distractions (Unestrcited environment test subjects only)

70% of the unrestricted test subjects performed multitasking while they were performing the usability testing; 64.3% of them reported that they were distracted by those additional tasks. In addition, 45% of the unrestricted test subjects experienced personal interruptions during the test. The test subjects' ratings to the effect of their distraction (multitasking and interruption) were analyzed; refer to Table 15 for the statistics.

TABLE XV. EXTERNAL DISTRACTIONS AND THEIR RATED EFFECT ON PERFORMANCE (RATING SCORES: [1] TO A VERY LARGE EXTENT – [5] TO A VERY SMALL EXTENT)

	(Average: STD), Sample Size	
	Multitasking	*Interruption*
Noumber of instances distractions	(1.78:1.1, N=9)	(2.13:1.13, N=8)
Rated effet of the distraction on perfomance	(3.67:0.9)	(4.1:0.9)

Fig. 3 shows the types of multitasking that have been experienced by the test subjects (x-axis) and their frequencies (y-axis). The same applies with Fig. 4 with respect to interruption types. However, some of those test subjects reported no distractions caused by 'multitasking' instances as they have not looked at them (they were not aware of), Fig. 3.

For interruption instances in the test, Fig. 4 clearly shows that 'type 2', (the SMSs and MMSs) were the most frequent cause of the test interruptions; they caused more distractions than those caused by phone calls in 'type 1', and they happened more times in a test session than the phone calls.

b) Types of Aparatus Used

The restricted test environment implies that the test subjects to use only UEA's machine in the UEA library's specified room utilizing the UEA's standard browser 'Safari' and UEA Network (NW). However, the test subjects in the unrestricted environment have used different machines, browsers, and NWs.

Of the unrestricted environment test subjects, 16 (80.0%) used their laptops; four test subjects used an Android phone, a notebook, a tablet, and a PC (1 test subject per each machine, 50.0%).

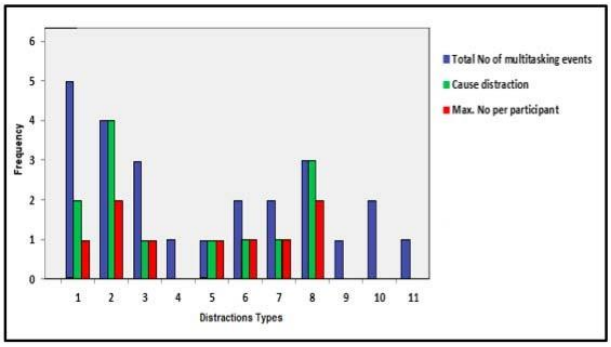

Fig. 3. Frequency of multitasking instances occurrence, ones cause distractions, and their maximum number of occurrence per the test session with regards to their type, types are: (1) Personal E-mail, (2) UEA webmail, (3) YouTube, (4) iTunes, (5) Chatting applications, (6) UEA Portal Website, (7) User's applications (e.g., Word Processor), (8) System popup messages, (9) Notes and demoes, (10) Other websites page opened in the same Internet browser's window (in another tab), and (11) Other websites opened in another Internet browser's window

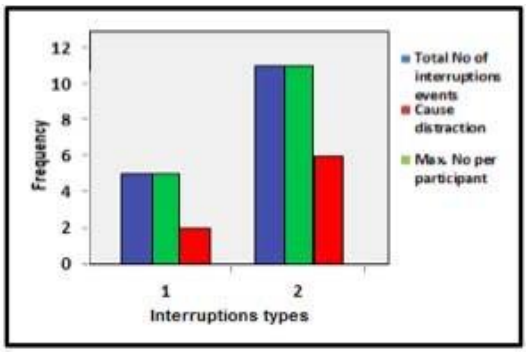

Fig. 4. Frequency of interruptions instances occurrence, ones cause distractions, and their maximum number of occurrence per test session with regards to their types

Sixteen test subjects in the unrestricted environment (80.0%) accessed the online usability testing website using WiFi technology via a DSL connection; one (5.0%) used his mobile to be connected (3G mobile connection technology) and three (15.0%) used a UEA network connection from their homes and offices. Thirteen (65.0%) test subjects used the Safari Web browser to access the online study website; five (25.0%) used Internet Explorer. The remaining two used Opera and Netscape (5.0% each).

Figure 5 shows a mapping between the machines, Web browsers, and NWs used in the unrestricted environment. The machines are represented by the bars and stacked by the type of network used (distinctively colored) and paneled by the type of browser used (the four partitions of the chart).

c) Mood at the Moment (Internal Distraction)

The test subjects were asked about their mood of the moment (while they were performing the test). By referring to the coded scores in Table 16 for the mood of the moment, we can see that that samples' mode and optimum scores are more positive in the restricted environment than in the unrestricted environment.

However, the Mann-Whitney test indicated that the optimum scores given in the restricted environment (Mdn = 3) do not differ significantly from ones in unrestricted environment (Mdn = 3), U = 58.500, z = -2.351, p =0.067, r =-0.3. The effect size r is considered to be small.

VI. Results Discussion

Online usability testing is a feasible method for usability testing in a non-lab environment. This preliminary conclusion is based on the finding that all participants who reported usability problems did so effectively, even without training on how to rate the problems before performing the test. Among the 99 existing problems, 42 were identified; therefore, 57 problems were agreed-on. In addition, all of the participants in the non-restricted environment who experienced distraction instances were able to report them. The data regarding the apparatus was collectable from test subjects in this environment using Loop11.

Most of the distracted participants were disrupted by multitasking as they attempted the test, but those interrupted personally were more affected by their interruptions. Hence, the test subjects often attempted multitasking while preforming the test in the non-restricted environment but were highly distracted by the personal interruptions to which they were exposed.

As described earlier in ref. [10], there was no significant difference in the time spent on tasks between the participants in the lab and natural environments. This pilot study found that there were also no valuable differences between the performances and the data gained between the online usability testing in the two testing environments. There was no significant difference in the time required to perform each task between the two test environments. This was also the case for the time for all of the tasks (the summation of the time per task). [6], [7] and [8] showed a significant difference in the time per task between the testing conducted in the lab and remote locations. However, they adopted different types of remote usability testing (synchronous and asynchronous).

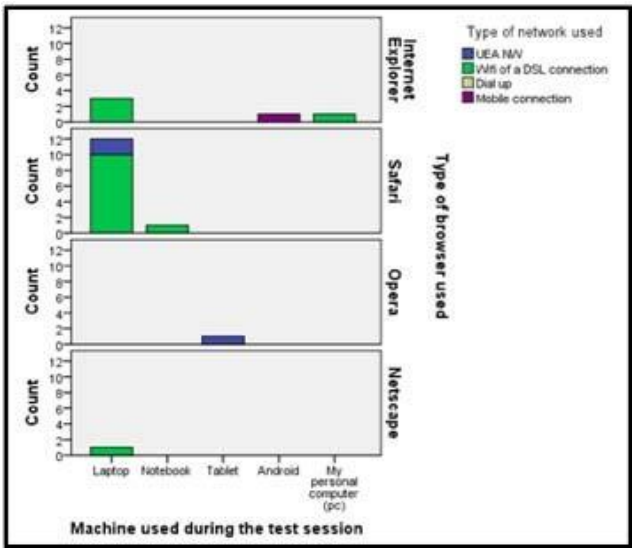

Fig. 5. Machines used in unrestricted environment stacked by type of NW utilized and paneled by type of browser used

TABLE XVI. Statistics of Mood of the Moment Ratings (Rating Scores: [1] Very Grumpy – [5] Very Happy)

Mood at the moment (1: Very happy – 5: Very grumpy)	**(Average: STD), Median**		**Mann-Whitney U test**
	Unrestricted environment	***Totally restricted environment***	
	(2.50:0.77), 3	(3.10:0.32), 3	(U=58.500, p=0.067, r=-0.3)

A significant difference was raised in the time required to complete the test questions. Accordingly, the time for the whole test (the sum of the time for all tasks and time per question) was affected. This increased significance was likely due the differences in the time per question that was included within the time for the whole test. This, in turn, could be because of the cognitive prescience to understand the English language by English non-native speakers. The questions were long and some participants reported difficulty understanding them on the first look. Of the restricted environment test' subjects, 60% were English non-native speakers; of them, 25% composed unrestricted environment test subjects. Ref [10] has also recorded the times per question and test, and the results of [10]'s study were similar to this study, as he also found a statistically significant difference in the time spent on questions between the two types of test environments. However, his study was administered in German, and 21% of the lab test's subjects were German non-native speakers; 13.5% of the natural environment test subjects were German non-native speakers.

There was no significance difference in the number of page views; yet, the effect size was considered small. This agreed with [10]'s findings. This also means that, even if participants spent more time performing the test, it was not necessarily because they had difficulty with the tasks. The reason for the longer time might have been caused by distractions, apparatus used, or any other factors.

Ref [10] was unable to conclude whether one environment yielded more efficient results than another.

The goal of online usability testing is to allow effective performance regardless of environment. The test subjects should complete a similar number of successful tasks and identify a similar number of usability problems in the test sessions. This was the case in a study by [6] and [7] wherein they applied remote synchrony usability testing. Their findings contradicted those of [8] and [9]. However, a greater number of problems were identified in the whole test administered in the unrestricted environment samples. The sample size in the unrestricted environment was larger than that of the restricted environment [6].

The test subjects in the unrestricted environment tended to identify more usability problems that agreed on the belonging-to categories in category and content. The restricted environment test's subjects tended to give more critical ratings than did unrestricted environment test subjects; that might have been a questionable determination since a large number of ratings in the unrestricted environment were unreported. Usability problems identified by unrestricted test subjects compose a greater percentage of problems that fall in any two-

agreement instances. Yet, the unrestricted test subjects tended to underreport the severity ratings of those problems. They also tended to identify more unique usability problems. These results are questionable because of the small sample size of the usability problems in each agreement instance, which makes the researcher unable to conduct a statistical analysis for significance.

The test subjects rated the task complexity, usability aspects of test objects, and satisfaction with the test location and setting similarly in the two environments.

VII. Lessons Learned, Limitations, and Future Work

Considering the lessons learned and the results gained from the pilot study, the researcher found it feasible to continue the research in the same direction toward investigating the effect of the distraction and non-standardization of the apparatus when performing online usability testing in different environments.

The work of the current formal study of which we are about to complete its design mainly addresses the issues and lessons experienced on the pilot study as follows:

- Screening questions should be adopted in the online usability testing to identify 'mental cheating' responses and to approach more relevant target users.
- To avoid the analysis overhead that was experienced using the online usability testing tools' judgments to measure task success, the test subjects should be asked a question after each task on the correct answer for that task.
- The extent of quantity and quality of required data to be collected from the online usability study is dependent on its design. Some participants indicated difficulty understanding the instructions, so the online usability study should be carefully designed to address all of the necessary information, and tasks and questions should be worded in clear, unbiased language.
- It should be made clear to participants that the usability problems are those experienced on the Web sites where they are carrying out the usability testing, not the problems related to study design.
- A more dynamic and simpler way of enabling the test subjects to report data should be adopted regarding the usability problems that they have encountered during the test, their severity ratings, and their frequency of occurrence during the test. The same applies for reporting distraction events.

This study collected only claimed data from test subjects. They are not automatically detected; as these types of tools do not have such capabilities and such things require costly tools, which contradicts the aim of using the totally remote means in remote testing (cost and time efficiency).

The distractions and apparatus should be formalized as environmental factors. The environment has been acknowledged as an individual attribute of usage context [13]. However, that is not necessarily applicable in every practical situation [14].

Contextual usage in usability testing should be defined carefully and include within it the most affective factors. To that point, we should address these factors in usability testing, and that should only be done if usability testing developers become aware of the importance of addressing such factors. That could be done by conducting such types of studies, given the fact that addressing these factors in online usability testing tools is not technically difficult. It is only a matter of being aware of such requirements.

In near future, the formal study will be applied considering the two different environments, the issue listed above, and a larger sample size, but with a deeper investigation of the environmental effect on usability testing performance. It will be applied with a different online usability tool to give more flexible options in designing questions and instructions, but will function similarly to Loop11.

VIII. Conclusion

This paper has presented the aim, methodology, and results of a pilot study of ongoing research that concerns the ability of online usability study to collect data about environmental factors, and the implications of attempting usability testing from different environments to test performance. In both environments involved in this study, there was no evaluator presence and consisted of twenty test subjects in the unrestricted environment (the subject's ordinary natural environment) and ten test subjects in totally restricted subjects (controlled environment similar to artificial labs) where they were recruited to perform predefined search tasks using the same online usability tool. This should guarantee a reliable way to have valid comparative results between the environments involved in the experiment.

In the restricted environment, the usability testing was performed with UEA students using standard UEA apparatus.

This study finds that online usability testing is a feasible method for usability testing in a non-lab environment as it enables the gathering of data about the usability problems and environmental factors (distraction and apparatus used). However, these data are claimed from the test subjects. This raises a need of stressing the awareness of online usability testing tools' developers through an understanding of the importance of collecting the usability problems from test subjects via such tools and addressing environmental factors. Increasing the level of communication between the academic community and such usability testing tools' developing community should enhance this.

The results of whether performing the test in different environments affected the performance show no valuable differences in most of the study's measurements, which agree with the findings of [10]. The test subjects were frequently multitasking while they performed the usability test in their unrestricted environment, but were highly distracted if they were personally interrupted.

We believe that we need a larger sample size to validate this finding and suggest a much deeper definition of environmental factors in the usability testing context, which will be adopted in the next formal version of this study.

References

[1] Nielsen, J. , “Usability inspection methods,”. CHI 1994, ACM Press, 377-378, 1994.

[2] Wolf, G., et al, “The role of laboratory experiments in HCI: Help, hindrance, or ho-hum?,” ACM SIGCHI Bulletin, 20(SI), 265-268, 1989.

[3] Winckler, M. A., Freitas, C. M., & de Lima, J. V., “Usability remote evaluation for www,” CHI EA 2000: the CHI 2000 Extended Abstracts on Human factors in Computing Systems, The Hague, The Netherlands, 1-6 April 2000, ACM, New York, USA, 131–132, 2000.

[4] Lewis, J., “Sample sizes for usability tests: mostly math, not magic,” Interactions, 13, 29-33, 2006.

[5] Alharbi, A., Galauert, J., & Mayhew, P, “ The Effect of Test Environment on Usability Testing,” The International Conferences on Interfaces and Human Computer Interaction 2014, Game and Entertainment Technologies 2014 and Computer Graphics, Visualization, Computer Vision and Image Processing 2014, Lisbon, Portugal, 15-17 July 2014, pp. 360-364. IADIS (In Press), 2014.

[6] Andreasen, S., Nielsen, V., Schrøder, O., & Stage, J. “ What happened to remote usability testing?,” An empirical study of three methods. CHI 2007: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems in San Jose, California, USA, 28 April - 3 May 2007. ACM, New York, USA, 1405-1414, 2007.

[7] Andrzejczak, C., and Liu, D. “The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience,” The Journal of Systems and Software, 83(7), 1258–1266, 2010.

[8] Bruun, A., Gull, P., Hofmeister, L., & Stage, J. “Let your users do the testing: A comparison of three remote asynchronous usability testing methods,” CHI 200: the SIGCH Conference on Human Factors in Computing Systems, Boston, USA, 4-9 April 2009. ACM, New York, USA, 1619–1628, 2007.

[9] Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. “An empirical comparison of lab and remote usability testing of Web sites,” In Proceeding of Usability Professionals Association Conference, Orlando, USA, July 2002.

[10] Greifeneder Elke, “Does it matter where we test? - Online user studies in digital libraries in natural environments,” Philosophische Fakultät I, 2011.

[11] Khanum, A., & Trivedi, C. “Comparison of Testing Environments with Children for Usability Problem Identification,” In International Journal of Engineering & Technology, 5(3), 2048-2053, 2013.

[12] Nielsen, J., and Molich, R. “Heuristic evaluation of user interfaces”, Proc. ACM CHI'90 Conf. (Seattle, WA, 249-256, 1990.

[13] Brown, P.J. The Stick-e Document: a Framework for Creating Context-Aware Applications. Elec- (tronic Publishing ’96 (1996) 259-272

[14] Abowd, Gregory D., et al. “Towards a better understanding of context and context-awareness," *Handheld and ubiquitous computing*. Springer Berlin Heidelberg, 1999