A Comparative Study of Synchronous and Asynchronous Remote Usability Testing Methods

Article · January 2013

CITATIONS
7

READS
988

4 authors:

Ahmed S. Alghamdi
Curtin University
8 PUBLICATIONS **55** CITATIONS
SEE PROFILE

Ali Al-Badi
Gulf College Oman
147 PUBLICATIONS **2,421** CITATIONS
SEE PROFILE

Roobaea Alroobaea
Taif University
273 PUBLICATIONS **3,642** CITATIONS
SEE PROFILE

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

www.irbas.academyirmbr.com
International Review of Basic and Applied Sciences
August 2013
Vol. 1 Issue.3

A Comparative Study of Synchronous and Asynchronous Remote Usability Testing Methods

AHMED S. ALGHAMDI
School of Computing Sciences, UEA, Norwich, UK
Email: as-123@hotmail.co.uk

ALI H. AL-BADI
Information Systems Dept., CEPS, SQU, Oman
Email: aalbadi@squ.edu.om

ROOBAEA ALROOBAEA
Faculty of Computer Science and Information Technology Taif University, Saudi Arabia
Email: r.alrobaea@uea.ac.uk

PAM J. MAYHEW
School of Computing Sciences, UEA, Norwich, UK
Email: p.mayhew@uea.ac.uk

Abstract

Traditional in-lab usability testing has long been the standard method for evaluating and improving the usability of software interfaces. 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. One of the most commonly used is the remote usability testing method. Ever since remote usability testing was introduced fourteen years ago, its effectiveness has been judged and evaluated against that of traditional in-lab testing. However, there is a distinct lack of research exploring the effectiveness of the various modes of remote usability testing. This research aimed to conduct a comparative study of two types of remote usability testing methods namely: synchronous and asynchronous remote usability testing. These two methods were compared through an evaluation of a website, which involved three points of comparison: number and type of problems discovered, overall task performance, and test participants' satisfaction. The results of the study showed that the synchronous testing method performed better than the asynchronous testing method in terms of the number and types of usability problems discovered, although no statistical significant differences were found. The participants in the synchronous test were notably more successful than the participants in the asynchronous test in completing the test tasks. However, the asynchronous test participants were significantly quicker than synchronous participants in performing those tasks. Participants in the synchronous tests also scored slightly higher satisfaction rate with regards to the targeted website. However, asynchronous participants were considerably more satisfied with the remote method that they had participated in. The paper concludes with a set of recommendations for conducting such research.

Keywords*: Usability testing, Remote usability testing, Synchronous and Asynchronous remote usability testing.*

Introduction

In today's dynamic software industry, usability is acknowledged as a fundamental quality factor. Many researches have illustrated the advantages of taking a usability centered approach during the software development life cycle. Amongst the numerous advantages cited regarding interfaces with high usability, are high end-user productivity and execution, and security [Scholtz, 2004].

To ascertain the usability level of a software system, various usability evaluation methods (UEMs) have been put forward in the last thirty years. Amongst these methods is the traditional in-lab testing method, which is considered the main method for assessing and improving the usability levels of software interfaces. Even though traditional in-lab testing methods are widely implemented, they are not without limitations, such as the lack of representative end-users, the expense of testing, and the lack of a proper simulation of the user's true environment [Hartson and Castillo, 1998].

To combat such limitations, other options and less costly UEMs have been put forward over the last two decades. One such UEM is the remote usability testing method. This method tackles the aforementioned limitations by employing actual end-users, and testing real scenarios in the end-user's true and natural environment. Remote usability testing is usually categorized into synchronous (moderated) and asynchronous (un-moderated) testing.

Synchronous remote usability testing is operated in real time, with the evaluator being spatially detached from the participants. Asynchronous remote testing is done by detaching the evaluators both temporally and spatially from the participants [Brush et al., 2004]. Ever since remote usability testing was introduced fourteen years ago, its effectiveness has been judged and evaluated against that of traditional in-lab testing [Andreasen et al., 2007]. However, there is a lack of research exploring the effectiveness of the various modes of remote testing methods (i.e. synchronous and asynchronous) in usability studies. This study endeavours to gain insight into this area.

The main aim of this study is to assess the effectiveness of the synchronous remote usability testing method against the asynchronous remote usability testing method when undertaking usability studies. The lower level objectives are: 1) explore remote usability testing methods and techniques; 2) apply synchronous and asynchronous remote usability testing to a targeted website; 3) compare synchronous and asynchronous remote testing outputs, so that each method's performance can be judged and evaluated; 4) generate a list of recommendations for further research regarding synchronous and asynchronous remote testing.

The guiding research questions [Creswell, 2009] of the present study are as follows:

1. Do the two remote testing styles (synchronous and asynchronous) vary in relation to the number and type of usability problems they yield?
2. Do the two methods vary in relation to task performance?
3. Do the two methods vary in relation to participants' satisfaction levels?

The paper is organized into five main sections. Section 1 includes introduction, research aim, objectives and research questions. Section 2 provides a literature review of the existing body of work related to usability and usability evaluation methods and techniques. Section 3 discusses the methodology employed in this research and provides a summary of various data collection techniques used. Section 4 has the results analysis and discussion which provides an outline of the qualitative and quantitative results from both the testing alternatives. Finally, section 5 has the conclusions and recommendations. It assesses the suitability of the methodology and looks at the extent to which the aims and objectives have been satisfied, as well as identifying the shortcomings of the study and providing recommendations for further exploration.

Literature Review and Background

Usability

The term "usability", was popularised in 1990s, in replacement of the term 'user-friendly', and has been defined in the existing body of literature in a variety of ways [Folmer and Bosch, 2004]. Hillier, (2003) defines usability as the extent an end-user can interact with ease with a system, without consciously giving it too much thought [Hillier, 2003]. Seffah et al. (2006), however, refer to usability as part of a product which enables users to execute their tasks quickly and with ease [Seffah et al., 2006; Dumas, 1999]. The International Organization for Standardization (ISO) states that usability of a product is "the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency,

and satisfaction in a specified context of use". Al-Badi and Mayhew (2010) defined usability as follows "Usability is the quality that indicates to what extent it is easy for all users of a website to interact with it in performing the required task(s)" [Al-Badi and Mayhew, 2010].

Usability is usually linked to five key attributes: learnability, efficiency of use, memorability, minimal errors, and user satisfaction [Nielsen, 1993]. The type of application will dictate which attribute will be most critical. For example, if the software is not used on a regular basis, then it is vital that users can recall with ease the actions needed for the desired tasks. If time is a key need of the application, then efficiency will be essential as well as the avoidance of errors [Scholtz, 2004]. A brief summary of usability attributes that appear in a number of different standards and models are summarised in Table 1.

Standard /Model	Constantine & Lockwood et al [1999]	ISO 9241-11 [1998]	Schneiderman [1992]	Nielsen [1993]	Preece et al. [1994]	Shackel [1991]
Attributes	Learnability	Efficiency	Time to learn	Learnability	Learnability	Learnability
	Efficiency in use	-	Speed of Performance	Efficiency of Use	Throughput	Effectiveness
	Memorability	-	Retention over time	Memorability	-	Retention
	Reliability in use	-	Rate of errors	Errors/safety	Throughput	Errors
	User satisfaction	Satisfaction	Satisfaction	Satisfaction	Attitude	Attitude

Table 1: Usability attributes of various standards and models [Seffah et al., 2006]

Website usability is widely recognised as the most important requirement for user acceptance. This requirement is especially critical for some websites, for example, for e-commerce websites; a customer dissatisfied as a result of poor usability is likely to become a competitor's customer. On the other hand, the user population is expanding in age, expectations, information needs, tasks, and user abilities. Websites should accommodate all these variations over time. Jakob Nielsen puts this very neatly in the following two quotations:

1) "Usability rules the web. Simply stated, if the customer can't find a product, then he or she will not buy it." 2) "The web is the ultimate customer-empowering environment. He or she who clicks the mouse gets to decide everything. It is so easy to go elsewhere; all the competitors in the world are but a mouse click away" [Nielsen, 1999].
A usable website is one that provides all the necessary functions and possesses a clear format and layout offering the user quick and easy access to whatever they require: these are the fundamentals of "website usability". Conversely, poor website usability can have a negative impact on various aspects of an organization, often resulting in considerable financial cost to a company and the services it offers [Wild and Macredie, 2000]. The impact of poor usability on commercial websites can have serious consequences in a competitive environment [Osterbauer et al., 2000]. According to Nielsen, 50% of potential Internet sales are abandoned due to poor website usability [Nielsen, 2001]. Thus, Internet businesses could potentially double their collective sales if e-shops could achieve a better standard regarding the quality of user experience [Nielsen, 2001].

Usability Evaluation

In order that usability evaluation methods (UEMs) and their evaluation can be fully comprehended, one must understand usability evaluation. Koutsabasis et al. (2007) defined usability evaluation as the appraisal of a particular application's user interface, an interaction metaphor or method, or an input device, to determine its actual or likely usability [Koutsabasis et al., 2007]. Overall, usability evaluation can be split

up into two general approaches, which are formative evaluation and summative evaluation. Formative evaluation refers to evaluation undertaking throughout development stages to advance a system design, while summative evaluation refers to evaluation undertaken once development has been completed, in order to appraise and judge the systems design (absolute or comparative).
Usability evaluation method refers to any method or procedure used to undertake usability evaluation of a specific application's user interface to highlight usability trouble areas. Some UEMs provide additional output, such as usability problem reports, to categorise usability issues according to type, to map issues to causative features from within the systems design, or to recommend alternative design solutions [Hartson et al., 2001].

During the last thirty years, several research studies have been conducted which seek to try and overcome usability problems in software systems. What stemmed from such studies was the development of a wide variety of evaluation methods (see Figure 1) [Scholtz, 2004]. From these evaluation methods, the most commonly employed are usability testing methods, usability inspection methods, and model-based methods [Scholtz, 2004]. Below is a short description of the three UEMs.

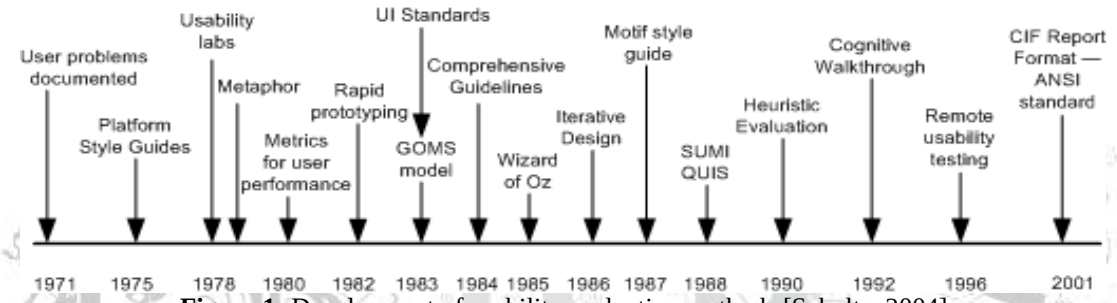

Figure 1: Development of usability evaluation methods [Scholtz, 2004]

- ***Usability Testing***

  Usability testing was the main evaluation framework used in the 1980s, and still holds the same precedence that it did when the concept was first introduced, specifically so for the latter stages of the design stages [Preece et al., 2002]. It enables the attainment of information about the way in which users use the system and pinpoints specific trouble spots within an interface. Many techniques exist for conducting usability testing, the most popular being, traditional in-lab testing and remote testing [Thompson et al., 2004].

- ***Usability Inspection***

  Usability inspection methods (otherwise known as expert-based evaluation) are a group of methods that involve having researchers (evaluators) monitor or investigate the usability aspect of a user interface. Such methods place emphasis on locating usability issues that might be met while using an interface, and providing suggestions improve the interface usability. There exist three main procedures for usability inspection methods: heuristic evaluation, cognitive walkthrough, and action analysis [Scholtz, 2004].

- ***Model-Based Evaluation***

  Model-based methods in usability evaluation are less popular than usability testing or usability inspection. They stem from work from the psychology field related to human performance. The main goal of adopting these methods is to forecast specific aspects of user performance on a certain interface, such as total task time or difficulty level of learning a tasks sequence. A good illustration of model-based methods is the GOMS model (stands for Goals, Operators, Methods and Selection Rules), which is used to forecast user performance with an interface and to forecast a task's time [Scholtz, 2004].

Concerns about Usability Evaluation Methods

There is no doubt that human-computer interaction (HCI) has advanced well with respect to utilizing evaluation methods to assess usability, but there are concerns regarding the optimal evaluation method. Despite this, many studies have sought to contrast such methods, however the comparison remain complex to test and many shortcomings have been identified with such studies. Firstly, is the matter of employing experimental (usability testing) techniques to obtain answers to general questions pertaining to usability instead of using more narrow questions that are more commonly utilized in experimental methods. Another matter is related to what measure should be utilised for comparison purposes. Should usability testing data be considered as being true and accurate? Not all usability tests are designed identically. They possess many shortcomings related to test design, testing manner, and analysis. However, despite specific methods being susceptible to limitations which can lead to a flawed implementation, there is no doubt that undertaking some evaluation methods is more beneficial than no action.

Presently, best practice involves employing several evaluation methods to obtain valid and reliable usability data. Evaluation techniques were essentially created to evaluate the usability levels of desktop systems. Thus the present focus in technological advancement of mobile and global computing brings with it new trials for existing usability evaluation techniques. In-lab traditional assessments will encounter difficulty in reproducing the natural use conditions for such applications. The necessity for direct field evaluation means that there are limitation on how early evaluations can be undertaken. Mobile and multi-user systems should be assessed for confidentiality and any usability matters pertaining to setting up, forming, and utilising such policies [Scholtz, 2004].

Nowadays, design and development life cycles in website creation occur at a rapid pace and proper usability evaluation is generally omitted in favour of shorter development times. Usability testers offer deeper consideration of remote testing methods, as they help to simulate the context of use for website evaluation in a more natural way. It is anticipated that innovative usability evaluation methods will be created and introduced to satisfy the requirements of our technology-orientated society. Investigators and practitioners alike will need to pool their expertise in order to solve such future usability challenges [Scholtz, 2004].

Remote Usability Testing

Remote usability testing is described as "usability evaluation where the test evaluators are separated in space and/or time from the test subjects" [Hartson et al., 1996]. The term "remote" indicates the distant location of the test participant from the evaluator (see Figure 2) [Castillo, 1997]. As mentioned earlier in section 1, remote usability evaluation can be separated into two key categories; synchronous and asynchronous methods.

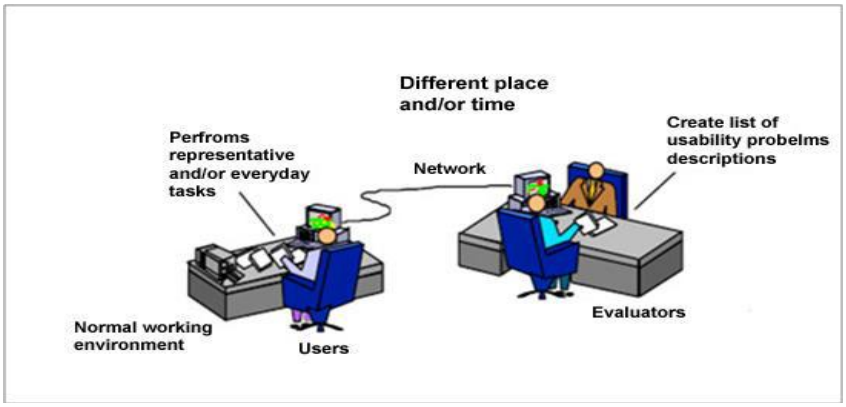

Figure 2: Remote usability testing [Castillo, 1997]

Synchronous Remote Testing:

Synchronous remote evaluation, also called 'live' or 'collaborative' remote evaluation, is a usability evaluation method that possesses a great deal of similarity to traditional in-lab usability evaluation [Selvaraj, 2004]. Such evaluation enables actual users to take part in the process from their natural environments, using their personal computers to keep the test conditions natural. In order to monitor the evaluation, the evaluator is located in the usability laboratory and his/her computer is linked to the remote participant's computer, via a real time Internet connection. This method also employs the use of video-conferencing or screen-sharing applications, and an audio connection via the computer or through a telephone. Such methods enable the evaluator to collate data on the user's activities by recording the whole scenario as the user undertakes the set tasks. The benefits of synchronous remote evaluation are the ability to obtain data from actual users in their normal environment, and reducing inconvenience for participants as there is no need for them to travel to a lab or test centre. However, limitations related to the Internet and telecommunications (poor bandwidth/delays) represent some of the disadvantages of this technique [Castillo, 1997].

Asynchronous Remote Testing:

Asynchronous testing splits the user from the evaluator in terms of location and time. Participants use their personal computers, which directs and measures user activities by the use of interactive programs or task based surveys [Winckler et al., 2000; Andreasen et al., 2007]. This method makes full use of the cost efficient and fast analysis that remote usability evaluation has to offer. Questionnaires are employed to guide the user, as opposed to making use of an evaluator, thus reducing financial and time resource pressures [Winckler et al., 2000; Andreasen et al., 2007]. Bastien (2008) asserts that the asynchronous approach is narrow in scope as it doesn't include observational data and recordings of sudden verbal data [Bastien, 2008]. This will limit the validity and accuracy of the results, which will lessen the chances of discovering usability problems. Conversely, due to cheap costs and easier user accessibility, this may potentially mean larger user sample sizes could be analysed, which would lead to far more accurate and realistic output. Greater sample sizes offer a truer representation of the users, while more natural test surroundings offsets the testing bias that may occur from a lab, which often leads to participants feeling pressurised which can affect the accuracy of usability results [Bastien, 2008].

Types of Remote Testing

Recently remote usability testing methods have received a great deal of scholarly attention, and numerous studies have been undertaken to further develop them. Consequently, a variety of different approaches exist, but overall three main techniques emerge from the literature, namely, instrumented remote evaluation, user-reported critical incident and remote questionnaire or survey [Petrie et al., 2006].

- **Instrumented Remote Evaluation:**

This technique refers to an automatic usability evaluation method in which applications are instrumented with an 'embedded code for capturing user data related to user actions and storing them in the form of logs' [Selvaraj, 2004]. These logs of data are assessed by evaluators using various pattern recognition techniques to help identify and pinpoint the nature of usability problem. The primary benefit of this approach is that user activity is not tampered with, they are not interrupted and there is no chance of interviewer bias. However, the downside of this technique is that it requires heavy human resources, so that the high volume of logs can be analysed [Castillo, 1997].

- **User-reported Critical Incident- Asynchronous:**

Previously known as the semi-instrumented remote evaluation method, is an asynchronous evaluation method developed by [Hartson et al. 1996] in 1996. Here, users are offered basic training to recognize

usage events which have noteworthy negative or positive effects on their task execution or acceptance, and then provide feedback about these events (e.g., problem outline and seriousness of the problem) while dealing with the application during their normal work routine. Reports are fed back to developers, alongside additional background information about the task and the system. Also information about screen-sequence activities is fed back to the developer as video footage. Evaluators then use this feedback to construct a list of usability problems for the user interface being examined. The disadvantages of this method pertain to the difficulty of training end-users and the total reliance on them to recognize critical incidents [Castillo, 1997].

- **Remote Questionnaire or Survey:**

This method entails giving questionnaire surveys to users (postal or via email or providing a URL link to a web-based questionnaire. The application under study can be supplemented to show a questionnaire which collates subjective data from users concerning the application and interface. The introduction of the questionnaire is "triggered by the occurrence of a specific event or a series of events during usage, such as the completion of a certain number of tasks" [Selvaraj, 2004]. A good illustration of this technique is the User Partnering (UP) module from UP Technology [Selvaraj, 2004], whereby specific events act as a trigger and prompt exchanges that request subjective user information concerning their practises of the application. Responses are then sent via the Internet to the developers for further analysis. This method is beneficial as it is able to extract the feedback of users whilst in their everyday environment, even though it is confined to new subjective data which stems from pre-programmed questions. Therefore, there is a distinct lack of qualitative data obtained during in-lab sessions during the use of this technique [Abelow, 1993].

How to Conduct a Usability Test

[Matera et al., 2006] asserted that there are six key stages which should be planned and organised carefully before undertaking an actual usability test, in order to minimise any inaccuracies pertaining to the reliability of results. These six stages are as follows:

Defining the Goals of the Test*:* Deciding on clear and concise goals of the test allows the evaluator to have a solid idea of what test outcomes are, and to outline suitable parameters for gauging that achievement. Usability testing goals can be very general in nature (e.g. measuring the user acceptance level of an interface) or very particular (e.g. assessing the readability of specific tabs on a navigation bar) [Matera et al., 2006].

Selecting Appropriate Data Collection Techniques*:* Currently, several methods are employed to collect data in usability testing sessions. Amongst these, the most popular are:

- ***Direct Observation****:* This method involves monitoring users' behavior whilst they engage with a website. Data is obtained from direct observations and can take into account objective information (e.g. time to finish tasks) and subjective information (e.g. users' annoyance or anxiety levels) [Benbunan-Fich, 2001].

- ***Thinking Aloud Protocol (TAP)****:* TAP is a type of direct observation whereby participants think loudly while undertaking a series of set tasks. This allows the evaluator to gain insight into how users see or understand the website interface under evaluation, and helps to recognize major issues that they may encounter when using the interface [George, 2005].

- ***Questionnaire****:* Questionnaires are beneficial for gauging the satisfaction levels of users and the simplicity with which they utilize an interface, and also for ascertaining users' attitudes. Using questionnaires is an indirect data collection technique, as they don't directly study the user interface, but rather place emphasis on users' opinions relating to the interface [Holzinger, 2005].

Selecting the Sample of Test Users: According to the existing literature, there are three major influences that must be taken into account before selecting participants for testing; number of participants, relevance of participants and experience of the participants. The reviewed literature shows that, there exists no set number of test participants, which are enough to uncover an appropriate number of usability issues. [Nielsen, 2000] has put forward that five users is sufficient to discern 85% of the usability problems (see Figure 3). [Turner et al., 2006] and others concur that the initial three to five users are able to identify the majority of usability issues, and that every extra test participant is less likely to reveal any fresh usability issues. Lewis (2006) however, asserted that a sizeable sample size, i.e. of up to 10 test participants, is necessary if the system has above average usability [Lewis, 2006]. There is almost a consensus amongst the evaluators that the test sample should be as truly representative as possible of the targeted users. Relevant users are more likely to encounter relevant problems which in turn will produce more relevant results. Experience of users may also be valuable, as experienced usability users are more likely to spot problems than beginners [Matera et al., 2006]. Potential criteria that can be used to ascertain the test sample may be the past user experience in usability evaluation (experts vs. novices), previous user experience with the application interface under assessment, and the user's age category [Matera et al., 2006].

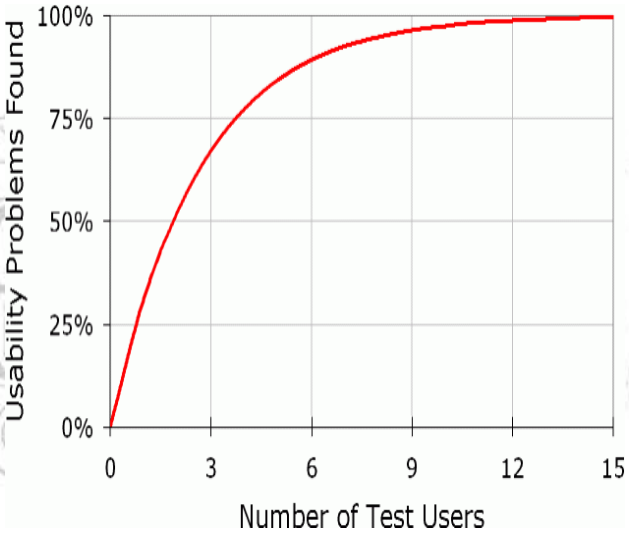

Figure 3: Curve showing relationship between problems found and number of users [Nielsen, 2000]

Designing Tasks: Usability testing tasks relate to activities that actual users would undertake when utilizing an application to achieve certain purposes, but they represent a key issue and a strong bearing in usability testing. There are various criteria for choosing and designing tasks such as task frequency and criticality. The former relates to tasks that are regularly utilised by users, whereas the latter pertains to the influence of tasks on a system's success [Macleod, 1994] . Other factors that have a bearing on task creation can be: Task generality, initial impressions, tasks which have innovative new features, edge-case tasks and tasks that the client/product team are concerned about. Task generality pertains to tasks that are usual; they aid in generalizing usability results after completing the usability test. Initial impression relates to tasks that can gauge user reactions in the first instance. Tasks with new features relate to tasks that can help in gauging the effect of any new features in a system. Edge-case tasks can highlight usability problems with big databases as well as other system usability issues. According to [Alshamari and Mayhew, 2008], usability testing tasks are usually grouped into three different categories: structured tasks, ambiguous tasks and problem-solving tasks. In structured tasks, users are directed stage by stage via the actions needed to meet their end goals. Ambiguous tasks however are built on the notion that there is uncertainty regarding whether or not users will locate the information that they require for while interacting with a website. Finally, in problem-solving tasks, users are given free-reign to behave naturally as they do in their everyday environment.

Establishing Usability Metrics: Prior to beginning a usability test, one must clearly outline what metrics will be employed to gauge a system's usability level. [Sauro and Kindlund, 2005] assert that the commonly employed usability metrics are:

- *Task completion rate:* This concerns the percentage of tasks that are completed correctly during usability testing.
- *Time spent on tasks:* As the name suggests, this measures the time it takes a user to perform a single task from start to completion.
- *Satisfaction score:* This identifies the typical user's level of satisfaction with an interface.

Preparing Test Material and Equipment: Prior to undertaking the experiment, the experimenter must make sure that all test material and equipment are available (e.g. instructions, surveys or observation sheets). It is also vital to undertake a pilot test run prior to the real test to ensure all test procedures are clear and modify them if necessary [Matera et al., 2006].

Rating the Severity of Usability Problems

According to [Desurvire, 1994] a usability problem can be described as "a problem encountered by the user and caused by an imperfect interaction". Measuring the criticality and more importantly, the extent of a usability problem, has been the aim of much research in the usability field. Severity and scope are the main aspects for gauging how pertinent a usability issue is in an evaluation and helps to ascertain what must be solved first in a product. This provides the development team with good insight and helps them to prioritize usability issues by the production of a structured list which highlights what must be modified prior to a product being released now and in the future. Various ways have been offered to categorise the severity of a usability problem. Amongst them, one popularly employed is Desurvire's Problem Severity Code (PSC) [Desurvire, 1994]. The PSC is a three level scale where a problem is categorised a "1: minor" which relates to minor anxiety or confusion. A problem rated a "2: serious" led to an error and a "3: critical" pertains to a problem that caused a task malfunction [Krista, 2006].

The Evaluator Effect in Usability Tests

The evaluator effect is a well-documented phenomenon which occurs in usability evaluations due to the type of usability metrics [Hertzum and Jacobsen, 2003]. Usability testing is usually undertaken once, whereby evaluators are under pressure to maximize testing of interfaces and create successful cost effective products with no flaws. The evaluator effect pertains to self trickery via the ideas, judgments and disagreements during usability tests. This happens when "the usability problems highlighted by one evaluator often shows little similarity to the sets offered by other evaluators evaluating the same interface" [Lewis, 2001]. This can be reduced by ensuring a mutual understanding of usability criteria and defined measurement methods. Success depends on making sure that the evaluator team agrees on mutual goals and this is vital to producing valid and successful usability evaluation output. For example, making sure all evaluators are in agreement on the difference between a major and minor problem by concisely outlining them prior to usability testing. This will ensure consistency and universal guidelines for measuring critical and minor problem during testing leading to consistent results. Synchronous testing gains both pros and cons with regard to the evaluator effect due to the non-direct interaction between the evaluator and the participant. The automation of hardware and software usually means there is a need for fewer evaluators. However, facial looks and reactions are more complex to record and can lead to vagueness when recording results via a computer screen or communication machine. Testing will see the effect transferred to the participants themselves as they become the evaluators during the testing, noting their own activities, experiences and responses during tasks and activities. Having many 'participant evaluators' create more opinions regarding problems and errors when they occur, due to the high variance of opinions during testing and can lead to an even greater evaluator effect. Using multiple choice answer formats in questionnaires may minimize this problem because their answer options are limited.

The Impact of Cultural Backgrounds on Usability Testing

Cultural dissimilarities can act as a serious obstacle during usability testing owing to issues such as group un-uniformity, appropriate environments, content and manner. [Henderson, 1996] defines culture by a "manifestation of the patterns of thinking and behaviours" which encompasses a wide range of groups ranging from geographical nationality to disability groups. For instance, undertaking usability tests on participants with poor vision, must take into consider their special needs, such as the need for larger text and screen readers.

In a national context, an example is the evaluation of western created website using non-western participants (e.g. Arabs), whereby aspects of language, attitude, content and even the contextual environment all differ and need to be accounted for. Such issues create interference and bias, which could result in inaccurate results, and thus culture needs to be taken into consideration when undertaking usability testing [McLoughlin, 1999]. Without a doubt culture will have a dominant influence in both the development and testing of development products [Wild and Henderson, 1997]. In order to reduce the likelihood of cultural variables influence, both the testing product and its cultural group must be adequately considered from an evaluator and participant viewpoint.

Research Methodology

This section offers an overview of the research methodology that was adopted and the data collection techniques that were employed in this study. It starts with an illustration of the main components of the research methodology, identifies the targeted website, and outlines the test tasks and participants. Following this, the section defines the criteria used to evaluate and compare the effectiveness of the two remote testing methods, discusses the material and equipments used in each remote testing method, and explains the experimental procedure. Finally, the section concludes with a brief summary of the study's pilot test and its results.

Research Methodology Overview

In this study, the experimental methodology was selected as the most effective method of answering the research questions and achieving the research aim and objectives outlined. Moreover, this study adopts the use of both quantitative measurement and qualitative assessment in data collection and analysis. The research methodology elements are illustrated in Figure 4.

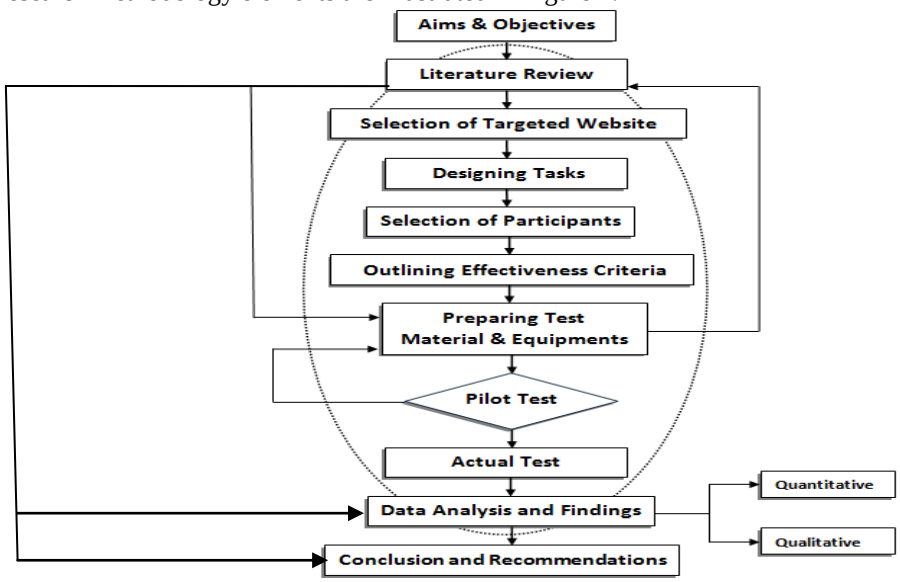

Figure 4: Main research methodology elements

Selection of the Targeted Website

Prior to undertaking this research, it was necessary to select a targeted website(s). Out of many websites, the website of the union of UEA students (UEAS) was selected to be to be the targeted website for this study (see Figure 5).

Figure 5: The targeted website

The rationale for using UEAS stems from two main reasons: the website of UEAS has an interactive interface with multiple functions, processes, and features. Also, the representative users of this website are easily accessible (as they are university students) and this will therefore assist and ease the sample selection procedure, providing representative participants who actively use this specific website.

Contact was established with the website administrators to obtain their consent to ensure in advance that there was no intention to modify or alter the interface during testing duration, and to gain preliminary information regarding the website for example "What tasks the users perform most commonly?" thus the researchers sought to obtain general data regarding user activities on the website, which would aid in the design stages of the test tasks.

Designing Test Tasks

Testing tasks in usability studies should be as representative as possible of the actual activities that end-users would perform whilst using the website. The researchers utilized the information obtained from the administrators of the website regarding user activity patterns to guide the design of various test tasks. Four tasks were designed and this was deemed an adequate number to ensure that the time that would be required to conduct the tests was not excessive. The tasks began by being relatively easy, and increased in complexity as information processing and associated cognitive demand became greater. All the tasks required information extraction but the level within the site hierarchy was variable which made the information somewhat challenging to locate. All tasks were deigned to be undertaken in isolation from one another, meaning that even if a task was not finished completely, participants would still be able to attempt the other tasks. This study employed problem-solving tasks, as these are the type of tasks that are most akin to the tasks that users commonly perform on the target website. Typically, the users of the targeted website are not offered guidance when undertaking activities; rather, they navigate the website independently and deal with problems as and when they are encountered.

The following four tasks were undertaken in both synchronous and asynchronous remote usability testing methods.

Task 1: Ascertain the opening times for the Union advice center

Suppose you want to obtain information regarding your academic studies from the student union advice center, but you are not aware of the weekday opening times of the center. Using the union of UEA student website, find out weekday opening hours of the union advice center.

Task 2: Locate the union election results

Suppose that your friend was a candidate for the role of women's officer in the recent elections, and you want to know the outcome. Using the union of UEA student website, find out who was elected for the post as the women's officer in the recent elections.

Task 3: Find the name of the financial officer

Suppose you are encountering financial difficulty and would like to seek advice from a UEA union officer, but are not aware of who to contact regarding financial issues. Using the union of UEA student website, find the name of the financial officer of the union council.

Task 4: Obtaining a tax exemption certificate

Suppose you are renting a house in the city and need to obtain a council tax exemption certificate from the university to prove your student status. Using the union of UEA student website, find out from where you can obtain a council tax exemption certificate.

Selection of Test Participants

The number, relevance and experience of potential participants are vital aspects in the sample selection process. There is debate regarding the preferable number of participants required in usability testing to obtain accurate results. Some researchers cite that eight participants are appropriate, while others state that ten participants are necessary for effective testing to occur [Lewis, 2006]. [Nielsen, 2006] contests that five participants are adequate for a sole usability test, but that a minimum of 20 participants are necessary if test results are to be contrasted with other tests or evaluation techniques, or if the results are used as part of a benchmark study. In this regard, a total of forty participants were recruited for this study, assigning twenty participants to each testing type (synchronous and asynchronous). In order to make the sample representative, the criterion for the sample selection was that all test participants must be university students as they are the target users of the chosen website. The method of sample recruitment was an explanatory email sent to students at the University of East Anglia, Norwich, UK, which invited them to participate in the study, as well as explaining the general purpose of the study.

	Synchronous Testing	**Asynchronous Testing**
Number of Test participants	20	20
Age	18-22 1 23-27 7 28+ 12	18-22 0 23-27 11 28+ 9
Gender	Male 14 Female 6	Male 13 Female 7
Country of origin	Saudi 10 Nigeria 4 UK 3 China 1 Turkey 1 Somali 1	Saudi 8 India 4 UK 2 China 4 Iraqi 1 Indonesian 1
Internet experience	100% use the Internet every day (average)	100% use the Internet every day (average)
Targeted website experience	13 participants experienced with the website	13 participants experienced with the website
Usability experience	5 participants experienced in usability issues	5 participants experienced in usability issues

Table 3: Demographic information of participants

Once participants were chosen, they were asked to fill out an online background pre-test questionnaire, which was developed using the online survey tool "SurveyMonkey" [SurveyMonkey, 2011]. The online questionnaire sought to obtain demographic information about the participants, such as their country of

origin, gender, experience regarding the Internet, prior experience with the targeted website, and previous experience regarding usability issues. The online questionnaire was completed in advance so that the sample could be fairly and evenly divided into the two remote testing methods. Table 3 provides an overview of the participants' demographic profiles for both the synchronous and asynchronous remote testing methods.

It should be noted that all test participants with prior experience in usability issues were students who have previously taken part in a course of "Human Computer Interaction" and completed a practical project in "Usability Testing".

Outlining the Effectiveness Criteria

To appraise the value of a usability testing method, and more importantly to compare the effectiveness of usability testing methods, the usability researcher must ascertain a definition for the comparison criteria. The criteria are stated in terms of one or more performance related indicators that are computed from raw empirical usability data (e.g., usability problem lists) yielded by each usability testing method. Deciding on the correct choice for criteria and performance indicators is dependent upon having a good understanding of the alternatives available and the shortcomings of each alternative. Thus, this stage concentrates on refining the criteria that would be employed to evaluate and compare the effectiveness of the two remote usability testing methods under examination. Therefore, an extensive literature review relating to studies of comparative usability testing methods was undertaken. The conclusions drawn from this review found that there is no singular standardised criterion for assessing and comparing the effectiveness of the usability testing methods, rather, usability researchers concentrate their research attention on different areas during the comparison of usability testing methods. Yet, the most popular criteria mentioned was regarding the number and types of usability problems discovered, overall task performance, and participants' satisfaction, as in the studies undertaken by [Andreasen et al., 2007] , [Brush et al., 2004] and [Thompson et al., 2004]. As these three specific criteria were utilised and referred to in the usability research; they were selected as the appropriate criteria framework to compare synchronous and asynchronous remote testing methods (see Figure 6).

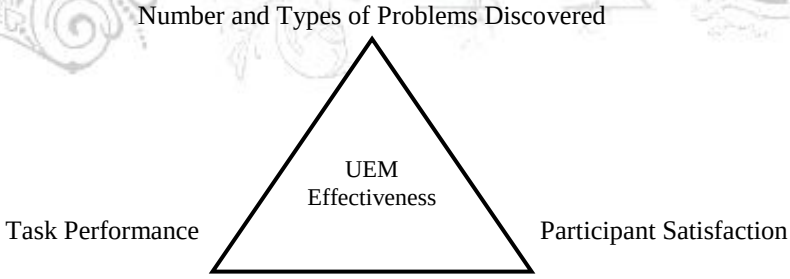

Figure 6: Effectiveness criteria

Number and Types of Problems Discovered

Four different indicators were identified in the usability literature to assess the number and nature of problems present in the two remote testing methods.

- ***Total number of usability problems discovered***: This indicator relates to the total number of problems found by each remote testing method. It has been used as a key measure of the effectiveness in usability testing techniques in various comparative studies such as [Andreasen et al., 2007] and [Brush et al., 2004].

- ***Types of usability problems discovered:*** As mentioned in section 2.5, [Desurvire, 1994] asserts that usability issues can be catergorised into several classes: minor problems, serious problems and critical problems. Evaluating the extent of usability issues uncovered by a specific usability testing methods

can aid in highlighting the extent of problems that method is able or unable to detect. This indicator has been the preferred performance measure in several comparative studies such as [Andreasen et al., 2007]. To ensure an objective assessment of the problems discovered, the final set of problems identified by the two testing methods was presented to a usability expert with over six years of experience in the field of usability engineering, who reviewed each problem in the set using the table 4.

Description	Score
Critical problem= caused a task failure/malfunction	3
Serious problem= caused an error	2
Minor problem=caused a minor anxiety or confusion	1

Table 4: Three point rating scale for severity of usability problems

- ***Number of unique usability problems uncovered****:* According to [Andreasen et al., 2007], unique problems are those identified by only one test method. It was the preferred indicator which was used in a comparative study of in-lab and synchronous remote testing by [Andreasen et al., 2007].

- ***Number of problems discovered in a test session****:* This indictor is concerned with the number of usability problems uncovered in a test session. It has been used in a number of comparative studies, such as [Andreasen et al., 2007].

Task Performance

The studies of [Thompson et al., 2004] and [Andreasen et al., 2007] have shown that task performance for testing methods has two main aspects: task completion rate and time spent on tasks.

- ***Task completion rate****:* Task completion rate relates to the percentage of tasks that are finished correctly during usability testing. The completion rate of tasks can be categorised into: either successful (completed) or unsuccessful (uncompleted).

- ***Time spent on tasks****:* The total time expended on each task, irrespective of whether the task was finished successfully, is noted. These timing can then be added together to calculate and identify the total time expended on all tasks.

Participant Satisfaction

Participant satisfaction has been employed as a key indicator in various comparative studies such as [Andreasen et al., 2007]. It was noted that various studies emphasised a range of aspects when testing for a participant's satisfaction. For instance, in Thompson et al. (2004) study, the emphasis was on obtaining subjective data from the participants regarding the degree of their satisfaction of the usability level of a targeted website [Thompson et al., 2004]. Whereas, the [Brush et al., 2004] study focused on gaining insight on the degree of participant satisfaction with the remote testing method they have used. For this research, the researchers devised an online post-test questionnaire survey using the Surveymonkey tool (Surveymonkey, 2011), which comprised of two main parts; the first part sought to gather subjective data relating to the degree of participant satisfaction with the targeted website, while the second section sought to gain information on their feelings regarding the actual testing method.

In order to collect data relating to the participants' satisfaction with the chosen website, the first part of the survey included both rating scale format and open-ended questions. The System Usability Scale (SUS) was utilized for the rating scale question. SUS, was founded by Brooke in the latter part of the 1980s and refers to a non-complex, ten-item scale which provides an overview of subjective measurements of usability [Brooke, 1996]. SUS is constructed using a likert scale question format and contains consistent content that leads to an inclusive usability and user satisfaction index (ranging from 0 to 100).

SUS is usually utilised after the user has interacted with the system under assessment. The researchers record the user's feedback each item. Once this is done, the researchers check all the items have been answered and tick the middle part of the scale when that user fails to respond to a certain item. Scoring SUS results in the generation of a single number which represents an overall measurement of the general usability of the system. To calculate the final SUS figure, all the scores from each item must be added up. Each item's score range goes from 0 to 4. For items 1, 3, 5, 7, and 9 the score contribution is the scale position minus 1. For items 2, 4, 6, 8 and 10, the score contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SUS. SUS scores have a range of 0 to 100 (Brooke, 1996).

In addition, participants were asked to answer open-ended questions on what they liked and disliked about the targeted website. The main purpose of these questions was to collect more qualitative feedback from the participants regarding their satisfaction of the chosen website and to understand the reasons for the ratings given for the rating scale question.

The second part of the survey collected opinions regarding the participant's satisfaction relating to the actual testing method itself. This comprised of questions in a rating scale format and open-ended questions. The rating scale format, has proved useful in research by [Brush et al., 2004] and [Selvaraj, 2004], who required respondents to rate the level of ease they felt when participating in the testing session, their comfort, their convenience, the degree to which they felt at ease when concentrating on the tasks, and their willingness to participate in a similar test in the future. A five-point rating scale was offered, where 1 represented "strongly disagree" and 5 represented "strongly agree".

Additionally, participants were asked to answer the open-ended questions about what they liked and disliked about the remote testing method in which they had taken part. The main purpose of these questions was to collect more qualitative feedback from the participants regarding their participation in the test session and to understand the reasons for the ratings given for the rating scale questions.

Preparing Test Material and Equipment

Synchronous Remote Testing

For the synchronous remote testing section of this study, the participants 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 browser, Skype Messenger application, and a connected functional microphone. Evaluation of the participants took place at the UEA Master's lab; the researchers used the same computer with all the participants in this section, which was equipped with Skype Messenger application, Internet browser, a microphone, and was connected to the participants through the Internet.

Skype Messenger 4.1 was used to connect and share the participant screens with the researchers so that the participant desktop could be observed for the duration of the test, and to communicate with the participant so that he/she could share their comments and suggestions. Skype was selected for this section of the study because of its availability, (Skype software is available for free download worldwide), and its simple, user-friendly interface [Skype, 2011]. Although most test participants had Skype on their computers, participants who did not have Skype were asked to download it from the Internet.

Asynchronous Remote Testing

The participants of asynchronous remote testing were required to have a computer with Internet connectivity and browser. They had to carry out the test without the help of any other equipment and within their own environment.

Loop11: Remote Testing Tool

To ascertain the usability of the targeted website remotely and generate appropriate solutions regarding the usability performance, the researchers used a web-based remote usability testing tool which is known as Loop11 [loop11, 2011]. Loop11 is a task driven tool used to collect data without the expensive physical

presence of a researcher in labs. By using Loop11, usability researchers can gain detailed insights and understandings of user behaviour. These insights are obtained via an interactive environment where users are asked to complete a series of tasks and questions on the interface of the targeted website. Interactions are captured, processed and made available in real-time reports. The data collected is both qualitative and quantitative applying metrics such as how many tasks have been successfully accomplished or how much time was spent on a particular task. Figure 7 shows two independent frames, which are opened when a participant starts evaluating the website with the help of Loop11. A small frame across the top of the screen presents an individual task being performed. The main frame, which fills the majority of the screen, is used to present the website under testing. Loop11 was employed in both remote testing methods.

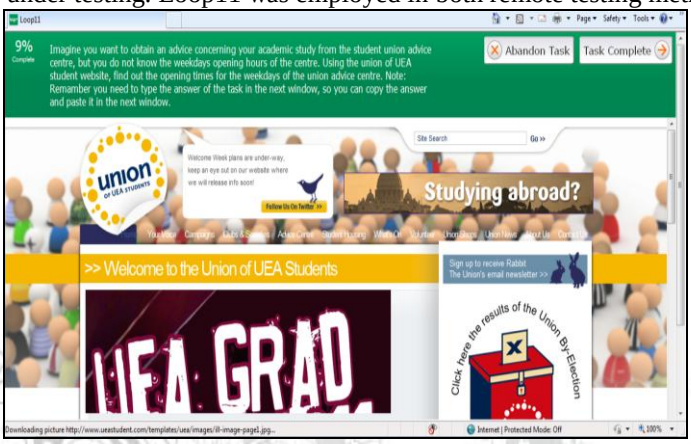

Figure 7: Loop11

Other Supporting Material

In both remote testing methods, supporting material was also prepared, such as a welcome page, informed consent form, online background questionnaire, online post-test questionnaire and an observation sheet.
Welcome page: Surveymonkey was used to provide a brief overview of the research and welcome the participants. The specific steps that needed to be followed were also explained in this document.
Informed consent form: The user's agreement is stated in this form that needs to be acknowledged by the participants prior to the initiation of the test. The online form was developed using SurveyMonkey. A hard copy of the informed consent form was also sent to the participants via email to be signed and returned.
Online background questionnaire: The information regarding the participants' background was collected using this questionnaire.
Online post-test questionnaire: This questionnaire was used to gather information concerning participants' perspectives regarding the targeted website and the testing session.
Observation sheet: The usability issues that were identified by a participant via synchronous remote testing were recorded by the researchers in an observation sheet.

Experimental Procedure

This study adopted similar procedures for both asynchronous and synchronous remote testing (see Figure 8). Each participant was assigned a number that would be used to identify their test information and files throughout the study. As the test was conducted online, every participant used their personal computers in their native environment. The test was not restricted to a particular location. An email was sent to the participants with the link of the welcome page and relevant information, such as the aims and objectives of the research. Rules, regulations, and procedures for the participants were also given on the welcome page. An online informed consent form needed to be approved once the participant had read and understood the brief introduction. Following that, an online background questionnaire was also needed to be completed. Then, participants were provided with a link to the target website and given 10 minutes to browse and familiarise with the targeted website's interface.

After that, participants were provided with a link to the test. Once the link of the test is opened, it directs participants towards the test and Loop11 opens two independent frames, as shown in Figure 7. Different queries were provided to a participant who then needs to browse the website to answer those questions. Participants in the synchronous group were encouraged to think aloud while performing the tasks. If a participant kept silent for a long time, he/she will be reminded by the researcher using Skype messenger. This enabled the researchers to identify issues regarding the website's usability. The problems that were identified by the participants in the synchronous group were also recorded by the researchers in an observation sheet. The asynchronous remote testing did not employ thinking aloud protocol.

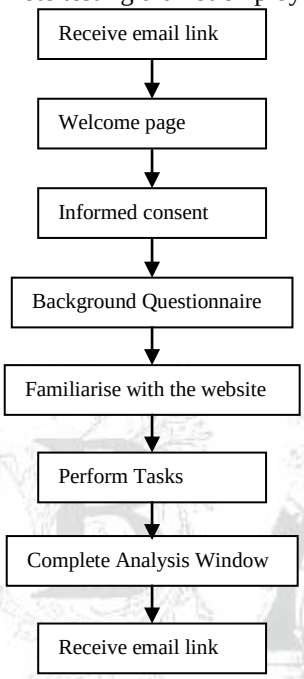

Figure 8: Experimental procedure

Once the answers were located the participants clicked on the "Task Complete" button located in the small frame. In the "Analysis Window", which was the next window; the participants were required to submit their answers by using the "Answer Field". The small frame also had the option of "Abandon Task", which was used by the participants when they were not able to answer the question. The option of "Did you encounter any usability problems? If so, please type them below" was used by the asynchronous group participants to record any usability issues that were faced while performing the task. This option was also available in the analysis window. The analysis window also had a "Next" button, to get to a new task. In this way the activities of participants were recorded, which included the time to perform a task, and the status (success or otherwise) of a completed task. Loop11 automatically recorded all the usability issues that were encountered by the participant, and all the tasks' answers that were input by each participant.

Once all the tasks were performed, an evaluation was required from the user to better understand the results. For this purpose, an online post-test questionnaire was developed. Additionally, as mentioned previously, participants were required to rate their satisfaction level with the remote testing method.

Pilot Test

Prior to undertaking the real test, a pilot was run to identify and fix any procedural problems such as inapplicable or unclear tasks [Matera et al., 2006]. The pilot test was run with participants who had no further involvement with the study. Two participants were chosen randomly from UEA students to take part in pilot tests for synchronous and asynchronous remote testing procedures. The pilot test revealed that the

tasks were not ordered correctly. According to [Nielsen, 1993], tasks should be ordered in ascending order of difficulty. However, the pilot test participants found the third task more difficult than the other tasks. As a result, this task was moved to the end of the test. All other aspects of the pilot test went smoothly and remained part of the actual test procedure.

Result Analysis and Discussion

This section starts by discussing the performance of the two remote usability methods in terms of the number and type of usability problems found. It then presents the results of test participants with regard to the task performance in both methods. Following that, the participants' satisfaction level with the targeted website and the test session is highlighted. Then, a statistical investigation of the correlations among the experimental variables is presented, before concluding with a brief summary.

Remote Testing Methods and Problems Discovered

This part focuses on analysing and discussing the number and types of problems identified by the two remote testing methods (RTMs). It is organised as follows: The first section (4.1.1) starts by discussing the total number of usability problems discovered by each method. The second section (4.1.2) discusses the types of usability problems discovered, the third section (4.1.3) discusses the number of unique problems discovered by each method, and the fourth section (4.1.4) discusses the number of problems discovered in a test session.

Number of Usability Problems Discovered

This section analyses and examines the number of usability issues discovered by the two remote testing groups. It is structured as follows: the first section highlights the total number of usability problems identified by the two remote testing methods. The second section presents the number of usability issues uncovered by male and female participants, while the third section discusses the number of problems discovered by expert participants in both groups. In the fourth section, the number of problems discovered by the best and first five users is examined. Lastly, the fifth section discusses the number of problem discovered during each testing task.

Total Number of Usability Problems Discovered

As mentioned previously in section 3.4, 40 participants were recruited for this study. The participants were divided equally into synchronous remote (SR) group and asynchronous remote (AR) group, with 20 participants in each group. An observation sheet was used to record usability problems in the synchronous group, whereas participants in the asynchronous group were asked to record usability issues in the analysis window after completing each task. The researchers added up the total number of problems encountered in each group, considering any repeated problems to be one problem, and then added up the total number of problems found by both remote testing methods. Table 5 illustrates the proportion and number of problems identified by each group and the total number of problems.

RTMs	Total Number of Problems Discovered	% of Problems Discovered
Total	21	100 %
SR	16	76.19 %
AR	13	61.90 %

Table 5: Numbers and percentages of problems discovered

Table 5 shows that both remote methods together were able to reveal a total of 21 usability problems on the targeted website. The SR group were slightly more effective than the AR group in identifying usability problems, as the SR group discovered 76% of the total number of usability problems found, whereas the

AR group identified almost 62% of the total number of problems uncovered. A possible explanation for this difference may be that in the SR testing sessions, the researchers were able to observe and communicate with the participants, who were also able to share their thoughts and comments while performing the test tasks applying a thinking aloud protocol. Accordingly, this can afford the opportunity for more usability problems to be noticed, whereas, in the AR testing session participants were required to record the usability problems themselves after completing the task. Another reason might be that SR participants were more concerned with how they would be judged by the researchers who were observing them and hence try to perform well in the evaluation and try making themselves look good by finding more usability problems. In general, these results are in line with the findings of [Andreasen et al., 2007], who concluded that SR testing outperformed AR testing in discovering usability problems.

Number of Problems Discovered by Male and Female Participants

The 20 participants in the SR group consisted of 14 males and six females, whereas the AR group consisted of 13 males and seven females. Since the majority of participants were male, it was unsurprising that they would find the largest number of usability problems in both groups (see figure 9). As mentioned previously, the total number of problems discovered by the SR group was 16. Male participants were able to reveal 14 usability issues of the total usability problems found by their group, whereas females were able to uncover 6 usability issues. When the usability problems found by males and females in the SR group were compared, it was found that 4 of the total usability problems were identified collectively. Thus, the males uniquely identified 10 usability issues while females identify two unique problems. It was observed that female participants in the SR group seemed to be reserved in communicating with the researchers and verbalising their thoughts and comments about the targeted websites, thus, minimizing the opportunity for more usability problems to be raised. This is perhaps because those female participants may feel shy and under pressure with a stranger watching, as one female participant stated, "*A stranger watching me while doing this makes me feel under pressure*".
On the other hand, in AR testing, male participants were able to identify 10 of the 13 problems discovered by their group, and female participants identified five. Again, this was expected as the female users were the minority in the AR test. In comparing the problems uncovered by the male and females participants in the AR group, it was revealed that two of the total usability issues were discovered commonly by males and females. Thus, the males uniquely identified eight usability issues while females identify three unique problems.

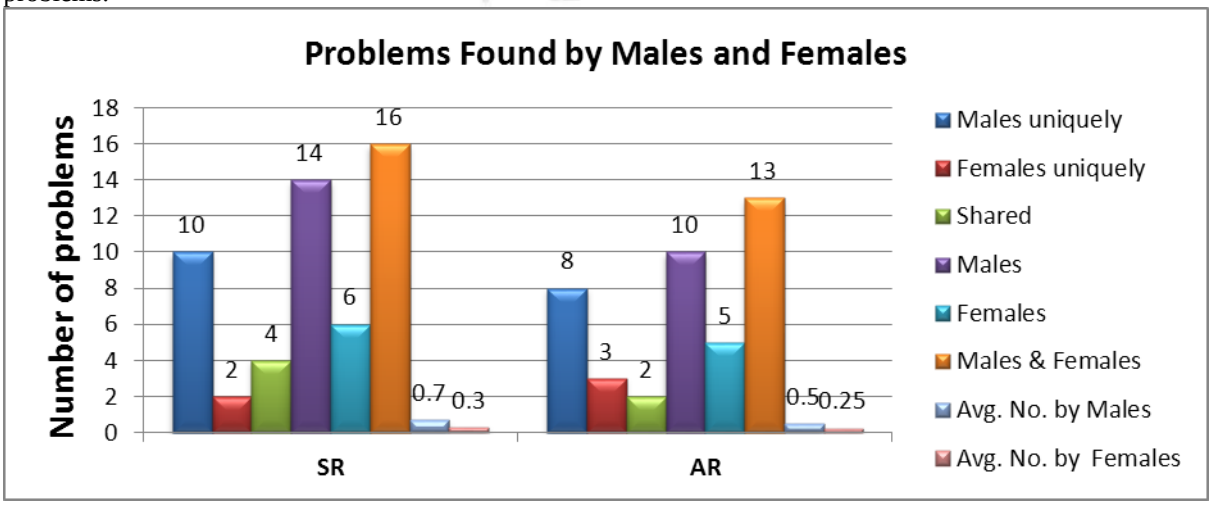

Figure 9: Number of problems found by male and female participants

Number of Problems Discovered by Expert Participants

The experience of participants in usability issues is a vital aspect in the sample selection process in usability testing. As mentioned previously in section 3.4, each group had five participants who were experts in the field of usability issues.

The following examines how these experts performed. The groups E-SR and E-AR consist of the expert five users (E) in usability issues for the SR group and AR group respectively. Table 6 shows that there are few differences in the characteristics of the expert users in both groups. The average for using the Internet every day of E-SR and E-AR is 100%, their ages ranged from 24 years to 27 years, and their experiences in usability issues are similar.

Groups	Gender	Age	Internet experience	Usability experience
E-SR	5 male 0 female	27.18 (average)	100% use the Internet every day (average)	all attended HCI course before
E-AR	4 male 1 female	24.31 (average)	100% use the Internet every day (average)	all attended HCI course before

Table 6: Details of the expert five users in groups E-SR and E-AR

Table 7 shows quantitative analyses for the expert five users in each group.

Group	Participant No.	No. Of Problems	% of Problems
E-SR	P 3	2	12.5%
	P 6	1	6.25%
	P 8	2	12.5%
	P17	1	6.25%
	P19	1	6.25%
	Shared	3	18.75%
	Total	**4**	**25.00 %**
E-AR	P 1	0	0.0 %
	P11	1	7.69 %
	P12	1	7.69 %
	P13	2	23.07 %
	P15	1	7.69 %
	Shared	3	23.07 %
	Total	**3**	**23.07 %**

Table 7: Expert five users (E) and number of problems discovered

It can be seen from table 7 that the E-SR group managed to reveal a total of 7 usability problems. Three of the 7 usability problems found were discovered by more than one expert in the E-SR group. Thus, the E-SR group only found a total of 4 problems out of those uncovered by their group. On the other hand, the E-AR group were able to identify five usability issues. Three of the five usability issues found were discovered by more than one expert in the E-AR group. Thus, the E-AR group were able to reveal 3 problems of the 13 that discovered by the E-AR group. The E-SR group and E-AR group discovered only 25% and 23% respectively of the total number of usability problems discovered by their respective group**.** Overall**,** it can be said that experience of users in usability issues in both groups did not play a big role in finding the usability issues.

Number of Problems Discovered by the Best and First Five Participants

Although it has been stated by a number of researchers that five users are enough to provide satisfactory results (i.e. finding 85% of usability problems, Nielson, 2000), in this research, using five users did not offer acceptable results. This section discusses this controversial argument and assesses different types of five users: the first and the best five users from each group, beginning with the best. The groups T-SR and T-AR consist of the top five users (T) who discovered the most problems for the SR group and AR group respectively.

Table 8 shows that there is little difference in the characteristics of the top performing five users in both groups. The average for using the Internet every day of T-SR and T-AR is 100% and their ages ranged from 25 years to 28 years.

Groups	Gender	Age	Internet experience
T-SR	4 male 1 female	28.23 (average)	100% use the Internet every day (average)
T-AR	3 male 2 female	25.18 (average)	100% use the Internet every day (average)

Table 8: Details of the top performing five users in groups T-SR and T-AR

Table 9 shows quantitative analyses for the top performing five users. The T-SR group and T-AR group discovered only 28% and 24% respectively of the total usability problems discovered, which is notably less than what has been claimed, i.e. 85%. The best results for five users could not reveal more than 28% of the total number of the usability problems found.

Top performing five users						(Nielson, 2000)		Maximum to be discovered	
Both groups		T-SR		T-AR					
#	%	#	%	#	%	#	%	#	%
6	28 %	6	28%	5	24%	18	85%	21	100%

Table 9: Top five users (T) and number of problems discovered (absolute and percentage of total number).

Figure 10 shows how the 20 users within each group performed. The proportion of usability problems found usually increased with the addition of each new user, although the rate of that increase was not large. Furthermore, it can be seen that the results of the first five users of the two groups did not exceed 16% of the usability problems found. The first five users from SR group performed better than the first five users in the AR group, but still notably less than the achieved results for both groups, which is 21 usability problems. Furthermore, Figure 10 also shows that the first 13 users from SR group and AR group just managed to find almost 50% of the found usability problems. Having discussed five users from different angles, such as the best performing and the first five performing users, it can be clearly seen that the "five users" argument is far from settled in usability testing in terms of achieving satisfactory results.

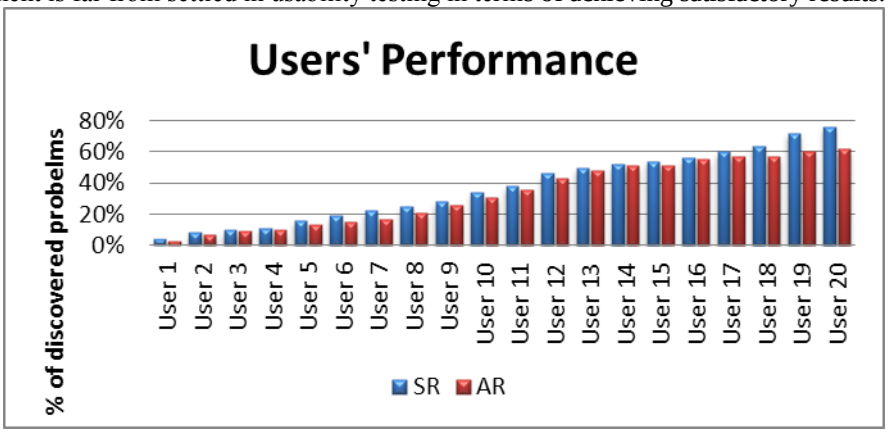

Figure 10: All users' performances in both groups (cumulative)

Number of Problems Discovered on each Task

The proportion of usability problems discovered by each task was determined in order to reveal the relationships between the difficulty of designed tasks and the proportion of usability problems discovered. It can be seen from figure 11 that for the SR group, the largest percentage of usability problems was identified on the forth task of the test (i.e. 61% of all usability problems found).

Similarly, 28% of all usability problems found by the AR group were identified on the fourth task. Conversely, the lowest proportion of usability problems in both groups was uncovered on the first task in the testing. These results are unsurprising, given that the tasks were in ascending order of difficulty.

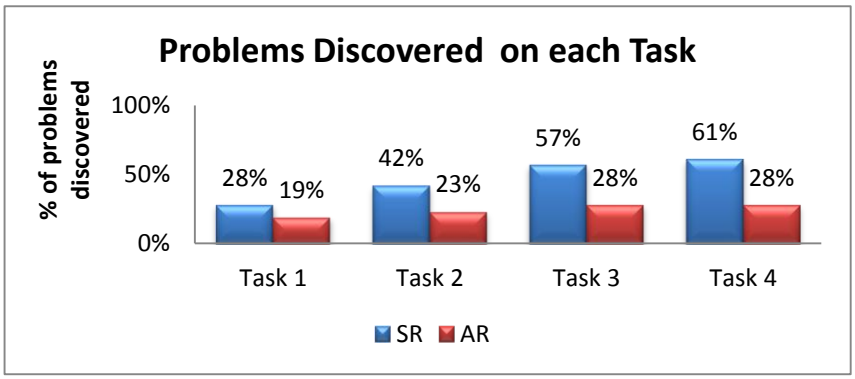

Figure 11: Number of problems discovered on each task

Types of Usability Problems Discovered

According to [Desurvire, 1994], usability problems can fall into one of the following categories of severity: minor, serious, and critical usability problems. At the end of both types of remote usability testing, the problems were collected and integrated into one complete list. Then, the final set of problems was sent to a usability expert with over six years of experience in order to ensure an objective assessment of the problems discovered in this study. The usability expert classified the severity of each problem as critical, serious or minor. Table 10 shows the characteristic of the usability expert involved in this classification process.

Gender	Country of Origin	Age	Degree	Job	Experience in Usability Engineering (Years)
Male	Oman	+ 30	PhD in Usability	Senior Lecturer	+ 6

Table 10: Profile of the usability expert

Figure 12 shows in detail which types of problems were discovered by each remote testing method. It can be seen from figure 12 that the 16 problems discovered by the SR group were classified by the usability expert into six minor problems, seven serious problems, and three critical problems. On the other hand, the 13 problems discovered by the AR group were classified into six minor problems, five serious problems and two critical problems.

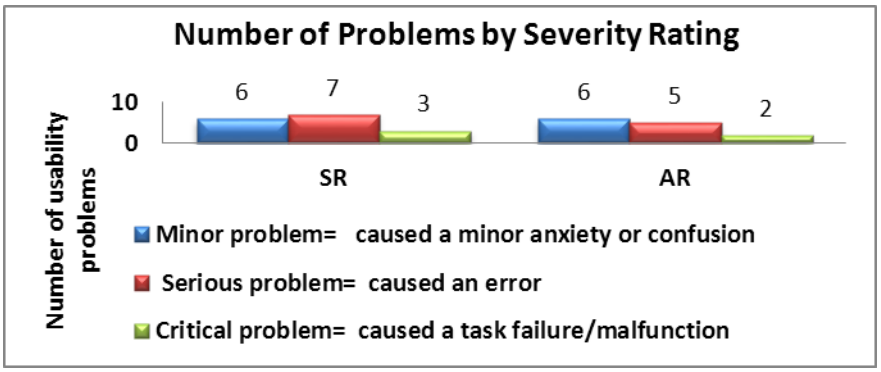

Figure 12: Number of problems by severity rating

Figure 12 shows minor and serious problems to be the two types of usability problems most identified by both the SR group and AR group, 6 minor problems identified by each group and 7 serious problems identified by SR and 5 by AR. Conversely, the critical problems represent the lowest number of problems discovered in both groups, 3 and 2 respectively. Interestingly, SR test and AR test revealed the same number of minor problems which was 6.

Figure 12 also revealed that the SR group performed better than the AR group in discovering serious and critical problems on the targeted website. As mentioned above, SR group managed to uncover 7 serious and 3 critical problems, whereas AR group were only able to identify 5 serious and 2 critical problems. These results are in line with the results of [Andreasen et al., 2007], which suggest that SR testing performed better than AR testing in detecting serious and critical problems. The Fisher exact test is a statistical test that examines whether the results of two or more small groups are statistically different [Weisstein, 2008] and [Sauro, 2006]. It can be applied here to examine the differences amongst groups' abilities to identify total, minor, serious, and critical problems. According to the Fisher exact test (see Table 11) there is no significant difference in the total number of problems identified in the two conditions (p=0.4801). In the identification of critical (p=1.000), serious (p=0.7311) and minor (p=1.000) problems no significant difference was found through the Fisher exact test.

	P
Total	0.4801
Critical	1.0000
Serious	0.7311
Minor	1.0000

Table 11: Fisher's exact test for the number of usability problems in each category identified in the SR and AR conditions, p=significance level.

Number of Unique Problems Discovered

According to [Andreasen et al., 2007], unique problems are those identified by only one test method. Table 12 below shows the remote test methods' performance on a unique basis i.e. unique performance, illustrating the number of problems revealed by each group but not identified by the other group. As shown in the table 12, the total number of problems discovered by the two groups was 21. SR group were able to reveal 16 usability issues of the total usability issues found, whereas AR group were able to uncover 13 usability issues. In comparing the two lists, the researchers found that 8 problems were on both lists. Thus, the SR group uniquely identified 8 usability issues while the AR group uniquely identified 5.

Table 12 shows that both groups collectively revealed 9 minor, 9 serious, and 3 critical usability problems. It also reveals that both groups identified the same number of minor problems, 6, but that each group identified three unique minor usability problems. Thus, they both performed at the same level of effectiveness in terms of revealing minor problems. SR group was also able to uncover four serious usability problems not identified by the AR group, whereas AR group identified uniquely two serious problems.

Problems Severity	Both Groups	SR Group	AR Group	SR Uniquely	AR Uniquely	Shared
Minor Problems	9	6	6	3	3	3
Serious Problems	9	7	5	4	2	3
Critical Problems	3	3	2	1	0	2
Total	21	16	13	8	5	8

Table 12: Number of unique and common problems discovered by each group

In addition, SR group individually managed to identify only one unique critical usability problem, whereas AR group was not able to identify any unique critical usability problem. However, two of the three critical usability problems were found on both groups' lists (a general terminology problem and unclear categories on the home page). In general, these results are in agreement with the findings of [Andreasen et al., 2007], which show that the SR testing method out-performed the AR testing method in identifying unique minor, serious, and critical usability problems.

Number of Problems Discovered in a Test Session

The number of usability problems identified in each test session (i.e. per participant) also varied between the two groups. Table 13 shows the average number of usability problems identified in each of the twenty test sessions that were conducted in each group. As seen from table 13, the average number of usability issues uncovered in SR group is slightly higher than that of AR group, with 2 and 1.05 problems respectively. This means that each SR participant was able to identify nearly double the usability issues that were uncovered by each AR participant. This result is consistent with the findings of [Andreasen et al., 2007].

RTMs	Average	SD
SR	2	1.654
AR	1.05	1.503

Table 13: Average number of usability problems identified in a test session

The number of usability problems found and their types are the most important measures in usability testing. The analysis of the two groups' performances in this study reveals a number of interesting results; it also helps to achieve the main research aim. So far, it can be seen that different types of remote usability testing methods indeed can yield different number and types of usability problems. It has been noticed that SR group were generally better than AR group in discovering usability problems. SR participants were also slightly more effective in uncovering more serious and critical usability problems, although no statistically significant differences were found. In addition, the SR test also outperformed the AR test in identifying unique serious and critical usability problems. Another interesting finding from this study was that the magic number of "five users" failed to provide satisfactory results (i.e. 85% of usability problems.) on the target website.

In the following sections, the task performance of the participants in both groups will be explored.

Remote Testing Methods and Task Performance

To measure task performance, the researchers used two indicators: the degree to which participants were successful in completing the test tasks and the time they spent working on them. The following sections analyse the results obtained from examining these two indicators.

Task Completion Rate

Task completion rate is a simple metric used to determine how well a site communicates with its intended audience. It concerns the percentage of tasks that are completed correctly during usability testing. Completion can be categorised into: either successful (completed) or unsuccessful (uncompleted). Tables 14 and 15 illustrate the performance of both the SR and AR groups. Each participant was asked to perform four tasks on the targeted website, meaning that a total of 80 tasks were performed by each group. Participants in the SR group successfully completed 66 tasks out of 80, whereas participants in the AR group were only able to complete 43. Participants in the SR test completed an average of 3.30 out of four tasks, in contrast to the average of 2.15 by participants in the AR group. This means that almost 50% of the tasks were not completed successfully in the AR group. A possible explanation for this difference might be that the psychological effect of the communication of the researchers with the SR test participants during the test might have given these participants more confidence when performing tasks.

Another reason might be that SR participants were more concerned with how they will be judged by the researchers who were observing them and hence try to concentrate more on the tasks. In general, these results are not in line with the findings of [Andreasen et al., 2007], which show that AR test out-performed SR test in performing the test tasks.

Tasks	Task 1	Task 2	Task 3	Task 4	Total	%	Average	SD
Successful	20	18	14	14	66	82.5 %	3.30	0.978
Unsuccessful	0	2	6	6	14	17.5 %	0.70	0.978
Total	20	20	20	20	80	100 %	4	0

Table 14: Number of tasks completed by SR group

Tasks	Task 1	Task 2	Task 3	Task 4	Total	%	Average	SD
Successful	15	12	10	6	43	53.75%	2.15	1.089
Unsuccessful	5	8	10	14	37	46.25%	1.85	1.089
Total	20	20	20	20	80	100%	4	0

Table 15: Number of tasks completed by AR group

Overall, in this study, task 4 was most difficult as half of the 40 participants were unable to complete it. In contrast, the first task was the simplest as 87% of the total number of participants was able to complete it. This was unsurprising, given that the tasks were assigned in ascending order of difficulty.
In the SR test only Task1 was completed by all participants, whereas no task was completed by all participants in the AR group. Figure 13 below compares the percentage of SR and AR participants who successfully completed each task.

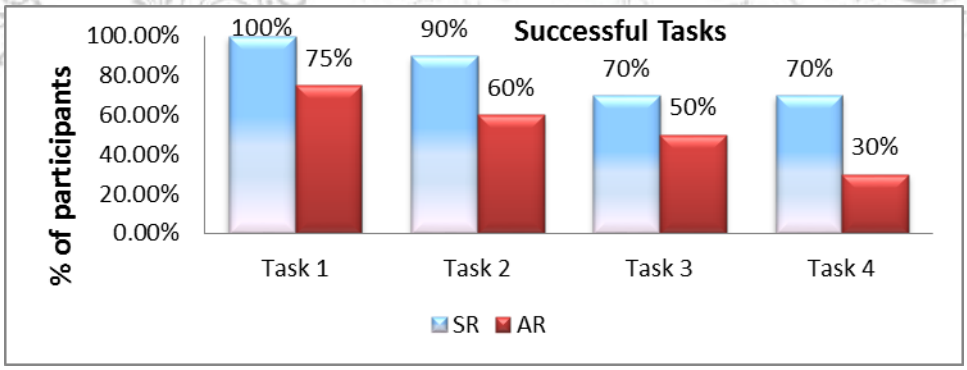

Figure 13: Percentage of participants who successfully completed each task

Time Spent on Tasks

This measure is aimed at revealing variations amongst the groups in terms of how long it took for the users to complete their tasks. Tables 16 and 17 compare the time spent by participants working on each task (regardless of whether the task was completed successfully), the total time spent on all tasks, the maximum and minimum time spent per task, the average time taken to perform each task and the standard deviation for each task. It is important to remember that the time spent on tasks by users includes both completed and uncompleted tasks. The uncompleted tasks sometimes needed less time than the completed ones as they gave up once they felt they were not able to perform it successfully.

Average Time/task	Task 1	Task 2	Task 3	Task 4	Total	Average	SD
Total Time (Secs)	2500	3359	3466	4223	13548	677.4	304.924
Maximum Time (Secs)	260	419	428	605			
Minimum Time (Secs)	42	48	19	19			

Table 16: Time spent by SR remote participants

Average Time/task	Task 1	Task 2	Task 3	Task 4	Total	Average	SD
Total Time (Secs)	2318	2777	2845	2926	10866	543.3	299.678
Maximum Time (Secs)	379	378	484	521			
Minimum Time (Secs)	11	14	13	30			

Table 17: Time spent by AR remote participants

Examining these results reveals that the participants in the SR group spent more time than the participants in the AR group, with 13548 and 10866 seconds respectively. This may be due to the fact that participants in the SR group had to think aloud while performing the tasks resulting in more time. The other reason might be that SR participants tried harder to complete the tasks to look good in front of the researchers who were monitoring them; as a result more time was spent. Overall, these results are in agreement with the findings of [Andreasen et al., 2007], which show that AR test participants complete tasks more quickly than SR participants.

The maximum time spent on a task was 605 seconds on the fourth task by one of the SR participants, while the minimum time spent on a task was 11 seconds on the first task by one of the AR participants. As expected, most of the time spent by users in both groups was on the fourth task, with 4223 seconds for SR test and 2926 seconds for AR test. Conversely, the amount of time spent on the first task was 2500 seconds for the SR group and 2318 seconds for the AR.
Figure 14 below compares the average time spent by each member on each individual task. The SR group took an average of 677.4 to perform their tasks, while the AR group took an average of 543.3.

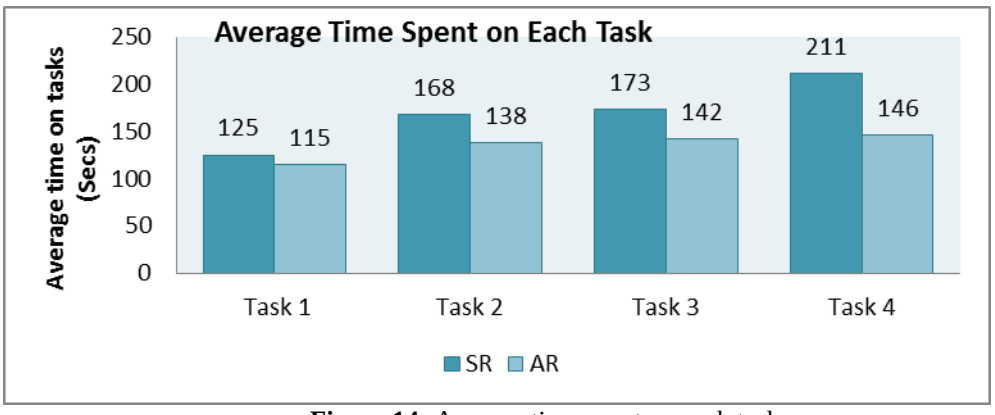

Figure 14: Average time spent on each task

From above analysis it can be said that the participants in the SR group were more successful than the participants in the AR group in completing the test tasks. However, the AR group were quicker than SR group in performing those tasks.

The next section discusses the participants' satisfaction level with regard to the targeted website and the remote usability testing method used.

Remote Testing Methods and Participant Satisfaction

As mentioned in section 3.5.3, an online post-test questionnaire consisting of rating scale questions and open-ended questions was used to understand the users' experience with the targeted website and the remote testing method they used. This section presents the results obtained from employing the online post-test questionnaire.

Participant Satisfaction and the Targeted Website

The first part of the online post-test questionnaire sought to gather data relating to the degree of participant satisfaction with the targeted website. It consisted of both rating scale questions and open-ended questions.

Rating scale questions

The System Usability Scale (SUS) was utilized by the researchers for the rating scale question. SUS consists of a ten-item scale which provides an overview of subjective measurements of the usability level of the targeted website. The best score is 100 and the worst score is 0. After calculating the averages of the SUS scores for SR and AR groups, it was found that the participants in the SR group scored a very slightly higher rate of satisfaction with the targeted website than participants in AR group, with 52.75 and 50.50, respectively. This may be because the number of successful tasks for the SR group was higher than for the AR group. Therefore, their satisfaction level was a little higher.

Open-ended questions

Each participant was asked the following open-ended questions:

- What do you like about this website?
- What do you dislike about this website?

The main purpose of these questions was to collect more qualitative feedback from the participants regarding their experience with the website and to understand the reasons for the ratings given for the rating scale questions. With regard to the first question, some SR participants stated that the website was simple, easy to use, and compatible to general standards. Others said that the website was dynamic and lively, with good navigation. Some of the AR participants stated that the website pages were attractive and the navigation through the website was easy. Additionally, four of the AR participants mentioned that the selection of font colour was reasonably good. With regard to the second question, SR participants highlighted a number of negative aspects of the website, such as, almost 72% of the SR participants stated that the font size was too small. Some of SR also said that there was a problem with the structure, navigation and terminology of the website. One of the participant said *that "some of the information was presented under inappropriate headings and so it was hard to find easily".* Nearly 82% of AR participants said that the font size was too small. Some of the participants stated that there was too much content on the homepage. Some also mentioned that there were too many pictures on the website. For instance, one participant said, *"too many flashing pictures which might disturb the website's visitor".*

Participant Satisfaction and the Remote Method Used

The second part of the online post-test questionnaire sought to gather data relating to the degree of participant satisfaction with the remote testing method used. It consisted of rating scale questions and open ended questions.

Rating scale questions

The rating scale questions, which were based on two earlier studies by [Selvaraj, 2004] and [Brush et al., 2004], asked participants to rate the following:

- The level of ease they felt when participating in the testing session.
- Their comfort when participating in the testing session.
- The degree to which it was convenient for them to participate in the testing session.
- The degree to which they felt at ease when concentrating on the tasks.
- Their willingness to participate in a similar test in the future.

The ratings were given on a five-point scale, with a rating of 1 for 'strongly disagree' and a rating of 5 for 'strongly agree'.

Figure 17 shows the participants' average ratings for each statement. The SR and AR groups gave almost similar ratings for the level of ease they felt when participating in the test session. However, the two groups gave different ratings for the comfort of participation in the study. The average rating given by SR group was 3.89, while the average rating given by the AR group was 4.27. This discrepancy was expected, as participants in the AR test did not need to download any software (i.e. Skype Messenger) onto their personal computer as the participants in the SR group were required to do. In addition, SR test participants were observed by the researchers while performing the test task. This perhaps made them feel uncomfortable.

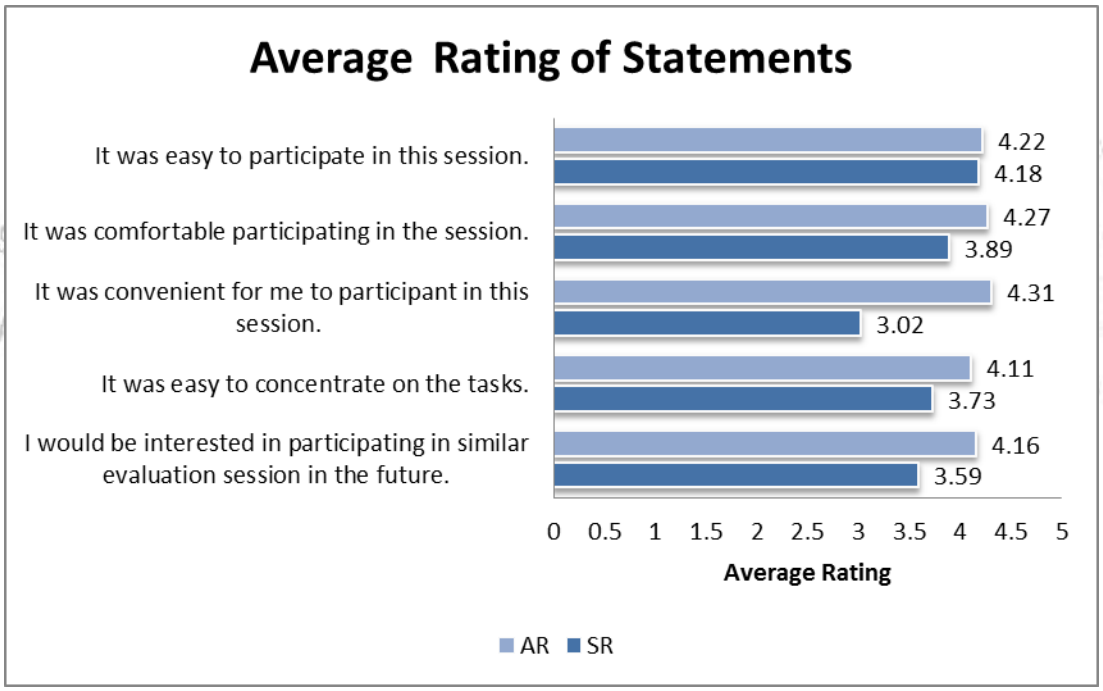

Figure 17: Average subjective ratings given by participants

Participants in the two groups also gave different ratings for the convenience of participation in the study. The average rating given by SR group was 3.02, while the average rating given by the AR group was 4.31. A possible explanation for this difference might be that AR test participants were able to do the experiments whenever they wanted, whereas SR participants had to schedule an appointment with the researchers to do the test session.

The statement about concentrating on the test tasks was also rated differently by the two groups. The average rating given by SR group was 3.73, while the average rating given by the AR group was 14. This was perhaps unsurprising, given that participants in SR group were required to think aloud while performing the test task, which might have distracted their concentration on the tasks, whereas AR test participants were not asked to do so.

The AR group showed more willingness to participate in a similar evaluation session in the future, giving this statement an average rating of 19. On the other hand, SR participants gave this statement an average rating of 3.59. This result was also expected, considering the average rating given by the two groups for the previous statements.

Open-ended questions

Each participant was asked to answer the following open-ended questions:

- What do you like about this remote usability testing method?
- What do you dislike about this remote usability testing method?

The main purpose of these questions was to collect more qualitative feedback from the participants regarding their participation in the test session and to understand the reasons for the ratings given for the rating scale questions.

With regard to the first question, almost 68% of SR test participants stated that the test session was simple, clear and easy to participate in, with common and realistic test tasks.
Most of the participants also mentioned that it is good to be able to do the experiments at your home or work place. Furthermore, some of the SR test participants also liked having the opportunity to communicate with the researchers and ask questions to clarify certain points while carrying out the test. For instance one participant said *"it makes it easier to communicate with the person concerned and get queries answered then and there so that the test tasks could be done as quickly as possible"*.

Most of the AR participants stated that it was very easy and convenient for them to participate in the test session as they could do so in their own environment and at their convenience. Additionally, 53% of the AR participants mentioned that the test tasks were clear and easy follow. One of the AR responses was *"this is a good way for the evaluation especially for those who have religious consideration that prevent them to share the evaluation with males"*.
With regard to the second question, some of the SR participants stated that they did not like being observed by the researchers while doing the experiment as they felt nervous and under pressure. Others said that being required to think aloud during the test distracted them from carrying out the tasks, such as, one of the SR test participants said *"I found it difficult to think aloud and perform the tasks at the same time"*. Moreover, some participants mentioned that they were quite concerned about downloading Skype Messenger on their computers for the purposes of screen sharing.

Most of AR test participants did not provide any negative feedback about the remote usability test method, apart from one who said that *"If I had a problem while doing the test, I'd have to try and work it out all by myself which could be frustrating"*.

Having explored and discussed the test participants' experience with the targeted website and the remote usability testing method used. It can be said that SR test participants were slightly more satisfied with targeted website than AR test participants. However, participants in AR group scored higher satisfaction level than SR group with regard to the remote method they have taken part in, showing more willingness to participate in similar evaluation session in the future.

Investigations of the Correlations amongst Variables

This section is designed to highlight any relationships between the experiment of variables: time spent, problems discovered, and user satisfaction with the targeted website (SUS). The Person Correlation Coefficient is used to reveal associations between the variables, and to identify the strength of any correlations found. It can also determine whether a correlation is positive or negative [Cohen, 2000]. [Cohen, 2000] offered a guideline for the interpretation of a correlation coefficient that explain the strength of correlation (none, small, medium and large) (see Table 18).

Correlation	Negative	Positive
None	−0.09 to 0.0	0.0 to 0.09
Weak	−0.3 to −0.1	0.1 to 0.3
Middle	−0.5 to −0.3	0.3 to 0.5
Strong	−1.0 to −0.5	0.5 to 1.0

Table 18: The Interpretation of a correlation coefficient [Cohen, 2000]

Table 19 shows the results obtained from employing the Pearson Correlation test.

Experiments Variables	Problem Discovered		Time Spent		User Satisfaction	
Groups	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**
Problem Discovered	1	1	0.63	0.53	0.23	0.17
Time Spent	0.63	0.53	1	1	0.15	0.09
User Satisfaction	0.23	0.17	0.15	0.09	1	1

Table 19: The Pearson correlations amongst the experiments' variables

Table 19 above shows the following results:

- There is a statistically strong positive relationship between time spent and problems discovered in both the SR and AR groups, where the p value is 0.63 and 0.53 respectively. This result reveals that the users who spent more time were able to discover more usability problems.
- There is no statistically strong relationship between time spent and user satisfaction with the website in both the SR and AR groups, where the p value is 0.15 and 0.09 respectively. This result reveals that despite the time spent on the website, user satisfaction was not influenced.
- There is no statistically strong relationship between problems discovered and user satisfaction in the SR and AR groups, where the p value is 0.23 and 0.17 respectively. These results reveal that the usability problems encountered by the users did not affect their satisfaction.

To sum up, it can be seen that the more time spent by a user, the more problems he/she found, but the time spent did not affect the users' satisfaction. Interestingly, identifying usability problems did not affect user satisfaction either.

Discussion

The results revealed that, overall, the SR testing method performed better than the AR testing method in terms of identifying usability problems, although no significant differences were found. As seen from table 20, the SR group were able to discover 16 out of the total usability problems found, whereas the AR group were able to find only 13. The SR test participants were also slightly more effective than AR test participants in uncovering more serious and critical usability problems on the website under testing. The SR group managed to uncover 7 serious and 3 critical problems, whereas AR group were only able to identify 5 serious and 2 critical problems. Furthermore, the SR test also outperformed the AR test in identifying unique usability problems, with 8 and 5 unique usability problems respectively.

	Usability Measures	SR	AR	Difference
1	Total number of problems discovered	16	13	3 (p=0.4801)
2	Number of critical problems discovered	3	2	1 (p=1.0000)
3	Number of serious problems discovered	7	5	2 (p=0.7311)
4	Number of minor problems discovered	6	6	0 (p=1.0000)
5	Number of unique problems discovered	8	5	3
6	Average tasks completed successfully out of 4	3.30	2.15	1.15
7	Average time spent on tasks per person (seconds)	677.4	543.3	134.1
8	Average SUS score	52.75	50.50	2.25

Table 20: SR and AR testing methods' performances, p=significance level.

Examining the results in table 20, it can be said that the participants in the SR group were notably more successful than the participants in the AR group in completing the test tasks. Nearly half of test tasks were not completed successfully, whereas, 82.5 % of tasks were completed successfully by SR test participants. However, the AR group were significantly quicker than SR group in performing the four test tasks. The SR test participants spent an average 134.1 seconds longer on the tasks than an AR test participant.

The statistical tests show that there is a strong correlation between time spent on a task and the number of usability problems found in both the SR and AR groups, where the p value is 0.63 and 0.53 respectively. This result reveals that the users who spent more time were able to discover more usability problems. Having discussed the performance of the expert five users in finding usability issues, it can be said that the experience of users in usability issues did not play a big role in finding usability issues in this study. Interestingly, the E-SR group and E-AR group discovered only 25% and 23% respectively of the total number of usability problems discovered by their respective group**.** Having explored five users' results through different categories (such as the best and first five users), it can be argued that the magic number of five users failed to achieve what it purports, i.e. identifying 85% of the total usability problems found: the best results for five users could not reveal more than 28%. Moreover, the results of the first five users of the two groups did not exceed 16% of the usability problems found. Thus, five users are not enough to provide satisfactory results.

With regard to participants' satisfaction with the targeted website, it can be seen from the results that the SR group scored a slightly higher rate of satisfaction with the targeted website than participants in the AR group, with 52.75 and 50.50, respectively.

With regard to the participant satisfaction with the remote testing method they participated in, the results revealed that, it was more comfortable, convenient, and easy to concentrate on tasks in the testing sessions for the AR test participants. The AR participants also show more willingness to take part in a similar test in the future than the SR test participants. Some of the SR test participants did not like having to think aloud while performing the test tasks as this distracted them from concentrating on the tasks. Also, some of participants in the SR group were quite concerned with downloading software onto their personal computers.

Conclusion and Recommendations

This section is divided into five sub-sections. The first section presents a conclusion of the research which provides answers to the research questions. The second section evaluates the degree to which the research aims and objectives were achieved, while the third suggests recommendations for researchers in this area. The fourth evaluates the research methods and tools used. The fifth section describes the limitations of the research and suggests possible future work.

Research Findings and Conclusion

The overall purpose of this study was to compare the effectiveness of the synchronous remote usability testing method against the asynchronous remote usability testing method when undertaking usability studies. This was measured according to the number and type of problems discovered, task performance, and the participants' satisfaction. The same website (*www.ueastudent.com*) was used for testing both methods.

The research raised three questions to be investigated. The first question was *"Do the two remote testing styles (synchronous and asynchronous remote testing) vary in relation to the number and type of usability problems they yield?".* As seen, the research revealed that, in general, the SR testing method performed slightly better than the AR testing method in terms of identifying more usability problems, although no statistical significant differences were found.

The SR test participants were also slightly more effective than AR test participants in uncovering more serious and critical usability problems. In addition, the SR test also outperformed the AR test in identifying more unique usability problems on the targeted website.

The second question was *"Do the two methods vary in relation to task performance? '*As now evident, the participants in the SR test were notably more successful than the participants in the AR group in completing the test tasks. However, the AR test participants were quicker than SR group in performing those tasks. Furthermore, it was found that there was a statistically strong positive relationship between time spent and problems discovered in both the SR and AR tests, meaning that the users who spent more time were able to discover more usability problems.

The third question was "*Do the two methods vary in relation to test participants' satisfaction with the process?*". The results indicate that the SR test participants were more satisfied with the targeted website; however, participants in the AR test were more satisfied than SR test participants with the remote usability testing method they had taken part in. It was more comfortable, convenient, and easy to concentrate on tasks in the testing sessions for the AR test participants. The AR participants also show more willingness to take part in a similar test in the future than SR test participants.

The main findings resulting from this study can be summarized as follows:

- The SR testing method performed slightly better than the AR testing method in terms of the number of usability problems discovered.
- The SR testing method outperformed the AR testing method in terms of the types of usability problems found. The SR method was able to reveal more serious, critical, and unique usability problems.
- Participants in the SR testing method completed more tasks successfully than the AR test participants. However, SR participants spent significantly more time on tasks.
- Participants in the SR test scored a slightly higher satisfaction rate with regard to the targeted website. However, AR participants were considerably more satisfied with the remote method they participated in.
- The statistical tests show that there is a correlation between time spent on a task and the number of usability problems found: users who spent more time revealed more problems. However, the tests did not show any relationships amongst the number of usability problems found and other variables such as user satisfaction.
- Having explored a large number of users' results through different categories (such as the best and first five users), it can be argued that the controversial number of five users failed to achieve what it purports, i.e. identifying 85% of the total usability problems found.
- Thinking aloud while performing the tasks distracted the SR participants from performing the tasks.
- Some SR participants were concerned with downloading any software their personal computers.

Evaluation of Research Aims and Objectives

This study sought to achieve one main aim and four lower level objectives. The following is an evaluation of the degree to which these were achieved.
The main aim of the study was to compare the effectiveness of the synchronous remote usability testing method against the asynchronous remote usability testing method when undertaking usability studies. This aim has been achieved, as seen in the preceding analyses.

The first objective of this study was to explore the concept of usability, usability testing methods and techniques. This was accomplished by reviewing many of the studies that have contributed to this research field, as shown in Section Two.
The second objective of this study was to apply synchronous testing and asynchronous remote usability testing to a targeted website. This was fully accomplished as described in Section Three.

The third objective of this study was to compare synchronous and asynchronous remote usability testing results in order to reveal each method's performance. This was accomplished by using figures and tables that make the comparison of both methods' performance easy and clear, as seen in Section Four. The final objective of this study was to produce a list of useful recommendations for future research regarding synchronous and asynchronous remote testing, which is accomplished in section 5.3.

Recommendations for Researchers

Based on the results of this study, participants' feedback, observations of the researchers and the researchers' experience in setting up the study, what follows is a list of recommendations for the benefit of researchers considering synchronous and asynchronous remote testing methods for usability evaluations.

1. Researchers should not require participants of synchronous remote test to think aloud during the test, as this distracts them from carrying out test tasks. Instead, it would more effective for participants to perform their tasks silently and comment on their work afterwards while watching a recording of their performance.
2. As some synchronous remote participants felt nervous and pressured, especially at the beginning of the session, researchers should set up a brief meeting with the synchronous remote participants before testing begins to remove pressure from the participants. Such a meeting also allows the researchers to ensure that the screen sharing application works and that the participant knows how to use it. Failure to do this may lead to connection problems that will frustrate both participants and researchers.
3. For synchronous remote tests, researchers should use screen sharing or web conferencing applications that do not require participants to download or install any major files, as users can be apprehensive of downloading extra software. In this study, some synchronous remote participants were concerned about downloading Skype Messenger.
4. Synchronous remote participants may not wish to share their computer screens if they consider their computers private. Hence, they should have remote access to researchers' computers and carry out test tasks in this way, rather than having their own computers or screens observed. This would also help companies to beta test products without distributing them.
5. Network connection speed plays a major role in most synchronous remote testing sessions; as a result, it is important to conduct remote tests over a high-speed Internet connection to minimise delay.
6. For asynchronous remote tests, researchers should provide test participants with guidelines to solve any problems that occur while performing the test. In addition, the researchers contact details should be provided to the participants.
7. Researchers should check with the website owner to find out whether they are planning to introduce any changes to the website interface, and/or should check the website daily to verify the stability of the interface.

Evaluation of Research Methods and Tools

Assessing the methodology used in this study can benefit future research in this area. The research methodology used in this study was the experimental method. It was an effective method of answering the research questions and achieving the research aims and objectives outlined. As mentioned in section 3.8, the data collection techniques used during the testing sessions were: the observation of participants' actions on-screen, thinking aloud protocol, questionnaires designed by Survey Monkey, and data collected via loop11 website. Observing the participants, while carrying out the test proved useful in gathering valuable information for the synchronous remote test. Once the screen sharing connection was established for the synchronous remote tests, it was easy for the researchers to observe participants during the test. In contrast, thinking aloud protocol proved to be an ineffective technique for collecting data. Participants in synchronous test were not comfortable with it, and many of them stated that thinking aloud while performing tasks distracted them from the tasks.

Questionnaires were used to gather data from participants in both remote testing methods regarding their experience with targeted website and the remote testing method used. The researchers used the Survey Monkey website to create and publish the questionnaires, which were then completed by the test participants [SurveyMonkey, 2011]. This was helpful for gathering data as all participants were located far from the researchers. The site collected and summarised participants' responses, which simplified the process of analysing the data.

The researchers also used the loop11 website to design and publish the task led/driven tests, and to collect the data from participants for both remote test conditions [loop11, 2011]. Loop11 was useful for gathering data. Part of this website's feature is that it summarizes the respondents' input which aids in data analysis and identifying trends. 40 participants were recruited for this study and divided equally into synchronous and asynchronous groups. All of the participants were students, since the dominant users of the targeted website fall into the student demographic. However, there was a gender imbalance among the participants; 27 were male and 13 were female. It would have been more effective if there had been an equal number of male and female participants. The test tasks were designed based on information gathered from the administrator of the target website. The researchers used the information gained to design tasks representative of the activities that the users of the website would normally perform. Participants' comments on the online post-test questionnaires expressed satisfaction with these tasks and stated that the tasks were very realistic.

Limitations and Future Research

Like any research, this study inevitably has its limitations but the researchers choose to view these not as weaknesses, but as opportunities for further research. As such, these possibilities can be divided into the following:

1. This study was based on a limited number of participants which might not be enough to generalize its findings. However, as future work, the whole study could be replicated on a greater number of participants in order to gain a better understanding.
2. The two types of remote usability testing methods used in this study (i.e. synchronous and asynchronous remote methods) were applied to one targeted website. Both types of remote usability testing methods could be applied to different websites such as commercial, or e-government websites, which may offer different and interesting results.
3. The performance of the two remote usability testing methods used was not judged against a benchmark usability testing method (e.g. traditional in-lab testing method). It would have been fruitful to compare the results of the two remote testing methods with a solid benchmark usability testing method.
4. As mentioned earlier, there were fewer female participants than male participants in each testing group. The research would have been more effective if the gender breakdown in the groups had been more balanced to reflect the university population.
5. The results of this study motivated the researchers to consider different types of remote usability testing methods, such as remote inspection or third-party laboratory evaluation. A possible fruitful avenue for future researchers may be to explore using new types of remote usability testing methods.

References

Abelow, D. (1993). Automating Feedback on Software Product Use. CASE Trends, 15-17.

Ali H. Al-Badi and Pam J. Mayhew, "A Framework for Designing Usable Localised Business Websites", Volume 2010 (2010), Article ID 184405, Communications of the IBIMA.

Alshamari, M. and Mayhew, P. (2008.) Task design: its impact on usability testing. *The Third International Conference on Internet and Web Applications and Services*. IEEE, 2008, 583–589.

Andreasen, M., Nielsen, H., Schroder, S., and Stage, J. (2007). What happened to remote usability testing?: an empirical study of three methods. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, page 1414. ACM.

Bastien, J. C. (2008). Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics, In Press, Corrected Proof.*

Benbunan-Fich, R. (2001). Using protocol analysis to evaluate the usability of a commercial web site. Information and Management, 39(2):151–163.

Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, pages 189-194.

Brush, A., Ames, M., and Davis, J. (2004). A comparison of synchronous remote and local usability studies for an expert interface. *In CHI'04 extended abstracts on Human factors in computing systems*, pages 1179–1182. ACM Press.

Castillo, J. (1997). The User-Reported Critical Incident Method for Remote Usability Evaluation.Master's thesis, Virginia Polytechnic Institute and State University, 1997.

Cohen, L., L. Manion& et al, 2000.Research Methods in Education. London, Routledge Falmer.

Creswell, J. (2009), *Research design: Qualitative, quantitative, and mixed methods approaches*, Sage Publications, Inc.

Desurvire, H. W. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In Nielsen, J. & Mack, R. (Eds.) *Usability inspection methods*. (pp.173-202). New York, NY: John Wiley and Sons.

Dumas, J. S. (1999).Practical Guide to Usability Testing. Chicago University Press, 2nd revised edition.

Folmer, E. and Bosch, J. (2004).Architecting for usability: a survey. *Journal of Systems and Software*, 70(1-2):61–78.

George, C. (2005). Usability testing and design of a library website: an iterative approach. OCLC Systems and Services, 21(3):167–180.

Hartson, H. R., Castillo, J. C., Kelso, J., and Neale, W. C. (1996). Remote evaluation: the network as an extension of the usability laboratory. In Proceedings of the SIGCHI conference on Human factors in computing systems: common ground, pages 228–235, Vancouver, British Columbia, Canada. ACM.

Hartson, H. and Castillo, J. (1998).Remote evaluation for post-deployment usability improvement. *In Proceedings of the Working Conference on Advanced Visual Interfaces* (AVI'98), May, (22-29). New York: ACM Press.

Hartson, H., Andre, T., and Williges, R. (2001).Criteria for evaluating usability evaluation methods.*International Journal of Human-Computer Interaction*, 13(4):373–410.

Hertzum, M. and Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 15(1):183204.

Hillier, M. (2003).The role of cultural context in multilingual website usability. *Electronic Commerce Research and Applications*, 2(1):2–14.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, 48(1):71–74.

Koutsabasis, P., Spyrou, T., and Darzentas, J. (2007).Evaluating usability evaluation methods: criteria, method and a case study. *In Proceedings of the 12th International Conference on Human-computer Interaction: interaction design and usability*, pages 569–578.Springer-Verlag.

Lewis, J. R. (2001). Introduction: Current issues in usability evaluation. International Journal of Human-Computer Interaction, 13(4):343.

Lewis, J. (2006). Sample sizes for usability tests: mostly math, not magic, Interactions, vol. 13, pp. 29–33, 2006.

Loop11 website. (2011). Task led/driven tool, available at: [http://www.loop11.com], (retrieved April, 2011).

Macleod M (1994) Usability in Context: Improving Quality of Use; in G Bradley and HW Hendricks (eds.) *Human Factors in Organizational Design and Management - IV - Proceedings of the I4th International Symposium on Human Factors in Organizational Design and Management,* (Stockholm, Sweden, May 29 - June 1 1994). Amsterdam, Elsevier / North Holland

Matera, M., Rizzo, F., and Carughi, G. (2006). Web Usability: Principles and Evaluation Methods. Web Engineering (Eds: Emilia Mendes and Nile Mosley), Springer.

McLoughlin, C. (1999). Culturally inclusive learning on the web, available at: [http://lsn.curtin.edu.au/tlf/tlf1999/mcloughlin.html], (retrieved March, 2011)

Nielsen, J. (1993). *Usability Engineering*, Academic Press, San Diego, CA, USA.

Nielsen, J. (1999), Designing Web Usability: The Practice of Simplicity, New Riders Publishing.

Nielsen, J. (2000). Why You Only Need to Test with 5 Users, available at: [http://www.useit.com/ alertbox/20000319.html**]**, (retrieved April, 2011)

Nielsen, J., (2001). Did Poor Usability Kill E-Commerce?, available at: [www.useit.com], (retrieved March, 2011).

Nielsen, J. (2006) Quantitative Studies: How Many Users to Test?,http://www.iadis.net/dl/final_uploads/200816C033.pdf, (retrieved May, 2011)

Osterbauer, C., M. K hle, T. Grechenig& M. Tscheligi, (2000). Web Usability Testing: Acase study of usability testing of chosen sites (banks, daily newspapers, insurances),in the Sixth Australian World Wide Web Conference.

Petrie, H., Hamilton, F., King, N., and Pavan, P. (2006).Remote usability evaluations with disabled people. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1141. ACM.

Preece, J., Rogers, Y., and Sharp, H. (2002*). Interaction Design Beyond Human Computer Interaction.* John Wiley and Sons Ltd.

Sauro, J. and Kindlund, E. (2005). Making Sense of Usability Metrics: Usability and Six Sigma. *Proceedings of the 14th Annual Conference of the Usability Professionals Association*.

Sauro, J., (2006). Sample Size Calculator for Discovering Problems in a User Interface, Measuring Usability

Scholtz, J. (2004). Usability evaluation. National Institute of Standards and Technology, *Encyclopedia of Human-Computer Interaction.*

Seffah, A., Donyaee, M., Kline, R., and Padda, H. (2006). Usability measurement and metrics: A consolidated model. *Software Quality Journal*, 14(2):159–178.

Selvaraj, P. (2004). Comparative Study of Synchronous Remote and Traditional In-Lab Usability Evaluation Methods. Masterarbeit, Virginia Polytechnic Institute and State University. Blacksburg.

Skype Messenger. (2011). Screen sharing software.http://www.skype.com/intl/en-gb/get-skype/on-your-computer/windows.(retrieved May, 2011)

SurveyMonkey website. (2011). A web-based survey solutions providers.http://www.surveymonkey.com. (retrieved June, 2011)

Thompson, K., Rozanski, E., and Haake, A. (2004). Here, there, anywhere: remote usability testing that works. *In Proceedings of the 5th Conference on Information Technology Education, pages 132–137.* ACM.

Turner, C., Lewis, J., and Nielsen, J. (2006). Determining usability test sample size. *International Encyclopedia of Ergonomics and Human Factors*, 3:3084–3088.

The UEA students union website. *http://* www.ueastudent.com, (retrieved March, 2011)

Weisstein, E., (2008). Fisher's Exact Test, http://mathworld.wolfram.com/FishersExactTest.html MathWorld (retrieved May, 2011)

Wild, M. and Henderson, L. (1997). Contextualizing learning in the world wide web: accounting for the impact of culture. Education and Information Technologies, 2(3):179–192.

Wild, P. & R. Macredie, (2000). Usability Evaluation and Interactive Systems Maintenance, *in Annual Conference for the Computer Human Interaction Australia.*

Winckler, M. A. A., Freitas, C. M. D. S., and de Lima, J. V. (2000).Usability remote evaluation for WWW. In CHI ’00 extended abstracts on Human factors in computing systems, pages 131–132, The Hague, The Netherlands. ACM.

Yin, R. (1994).*Case study research: design and methods, Applied Social Research Methods Series*, vol. 5. Thousand Oaks: Sage.