Implications of Applying Usability Testing with Remote Users

Abeer Abdullah K. Alharbi

Submitted for the degree of Doctor of Philosophy

School of Computing Sciences

University of East Anglia

December 2019

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

"وماتوفيقي الا بالله"

Without dreams, there can be no courage, and without courage, there can be no action.

— Wim Wenders

To those who have an ambition and determination,

those who believe in their selves,

those who never give up,

this thesis is dedicated for you

Abstract

Some studies in the literature on remote asynchronous usability testing have indicated the existence of contextual factors related to remote-uncontrolled environments. Typically, in these environments, users take part in the usability test at any time although uncontrolled contextual factors might be present. Moreover, such settings might induce different interactions with the evaluated products, which consequently may influence the data collected in the usability test. Therefore, this research aims to explore these kinds of interactions to determine whether they differ from users' interactions in the laboratory and, if so, how. The findings of this research are intended to contribute new knowledge about the implications of applying asynchronous usability testing to remote users. To meet this goal, three main studies are conducted: the first exploratory study is aimed at exploring what happens during testing sessions in users' natural environments. The second empirical study involves two participant samples: one sample performed the test in their natural environment, and the other sample performed the text in a lab. The performances of both groups are compared to explore their differences. User-reported data regarding contextual factors are also explored. In the third controlled experimental study, stimulating contextual factors are applied during usability testing sessions to explore the users' interactions.

The results showed that usability testing outcomes were independent of the method itself. With respect to physical environments, contextual factors were the most influential in the outcomes of usability testing. Although interruptions had the highest negative influence, the extent of this influence differed based on the type of interruption applied. In-person interruptions were the most disruptive because they influenced, not only the number of errors and task-load measurements, but also the time taken to perform tasks. Instant messaging increased the number of errors and the task load. Phone interruptions did not have noticeable effects on performance, but increased stress, time pressure and frustration. Based on our results, we concluded that if remote asynchronous usability testing is used, then the influence of contextual factors should be expected. Hence, these factors should be collected during testing because awareness of them is vital in improving data interpretation.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate ‘take down’ action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Acknowledgements

I would like to thank the Almighty Lord for giving me the strength and ability to complete this thesis. Some special people made this extraordinary journey much less complicated and more rewarding for me. First, I want to thank my supervisor, Dr. Pam Mayhew, who never wearied of reading the revised versions of this research. She always encouraged me each time I felt down. Her support and relentless insistence on quality, completeness and integrity have been invaluable. I also thank Prof. John Glauert (second former supervisor), Professorial Fellow, School of Computing Sciences, for the advice he gave me during the early period of this research. I also thank Dr. Joost Noppen (second supervisor) for his encouragement and insightful comments. In addition, I would like to express my deep gratitude to all the volunteers and anonymous students who participated in this research.

My personal gratitude is for my parents, Mr. Abdullah Alharbi and Mrs. Fedhah Alharbi, for the unconditional love and support that they have given me. Father and Mother, I thank you for always believing in me, praying for me and supporting me all the way. My husband, Dr. Sultan Alharbi, is my adviser and my very best friend. His encouragement has given me the motivation to finish something I dreamed of completing many years ago. With his incredible persistence, patience and courage, he did everything possible to help every step of the way.

My special thanks are for my little one, my daughter, Mayar, who suffered during my studies because she did not have the time she needed from me. Mayar, I cannot describe how much I love you. I thank my brothers, Mr. Mohammed, Mr. Khaled and Mr. Majed and to my only sister, Dr. Dalia, all of whom encouraged me by their prayers, moral support and never-ending love. I especially thank my older brother Mohammed and my sister Dalia for their great help during my difficult days. Without my family, this research would not have been completed.

I acknowledge my sponsors, Imam Mohammed Ibn at Saud University and the Ministry of Education in Saudi Arabia for recommending me to the Saudi Cultural Bureau and the Royal Embassy in the UK for the full scholarship awarded to me.

I could write several pages naming each family member, friend and colleague who contributed to my emotional and spiritual well-being during my PhD studies. However, I will just say thank you to everyone! God bless you all!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

4FFCF	Four-Factor Framework of Contextual Fidelity
B	Baseline
NE	Natural Environment
HCI	Human Computer Interaction
IM	Instant Messaging
NW	Network
OUUT	Online Unmoderated Usability Tool
Ph	Phone interruption
Pr	In-Person interruption
RAUT	Remote Asynchronous Usability Testing
SMEQ	Subjective Mental Effort Questionnaire
SMQ	The Single Ease Question
SUS	System Usability Scale
TAP	Think Aloud Protocol
TLX	NASA Task Load Index
UCI	Users Critical Incidents
UEM	Usability Evaluation Methods

List of Publications

The research work presented in this thesis is original work of my own, unless otherwise indicated in the text. Parts of this thesis have been published and/or presented at the following conferences:

1. ALHARBI, A., SMITH, D., & MAYHEW, P. (2013, OCTOBER). Web searching behaviour for academic resources. In *Science and Information Conference (SAI, 2013),* London, (pp. 104-113). IEEE.
2. ALHARBI, A., & MAYHEW, P. (2014, FEBRUARY). The Effect of Test Location and Environment on Usability Testing. 7th Saudi International conference. Edinburgh. ISBN: 9780956904522
3. ALHARBI, A., GLAUERT, J., MAYHEW, P. (2014, JULY). The effect of test environment on usability testing. In Information and Human Computer Interaction Conference (IHCI, 2014), Libson, Portugal, (pp. 360-364). In IADIS, ISBN: 9789898533227.
4. ALHARBI, A., & MAYHEW, P. (2015, JANUARY). User Environments' Implications for Usability Testing Performance. 8th Saudi International conference. London.
5. ALHARBI, A. & MAYHEW, P. (2015, SEPTEMBER). Users' performance in lab and non-lab enviornments through online usability testing: A case of evaluating the usability of digital academic libraries' websites, In Science and Information Conference (SAI, 2015), London, (pp. 151-161). IEEE. doi: 10.1109/SAI.2015.7237139.

Chapter 1: Introduction

1.1 Overview

User-based testing has become a de facto standard in usability engineering. The test assesses the usability of a system in a controlled laboratory environment where users are observed while interacting with the product. However, in some situations, it is neither possible nor preferable to apply usability testing to users in a laboratory. Some software organisations do not deploy systematic usability activities in their development process, and it would be a resource overhead for them to apply usability testing in a laboratory. For example, it is difficult for some software organisations that develop and evaluate products for global markets or practice outsourcing to apply usability testing when their developers, evaluators and users are distributed across software organisations, countries and time zones. Recruiting target users for global products, especially for websites such as e-commerce and digital library websites, is difficult and costly in terms of the time and effort required in a laboratory. In such situations, it is relevant to apply remote asynchronous usability testing (RAUT), which is the method used to overcome the drawback of resource overheads. RAUT enables increased access to participants and reduces travel expenses.

RAUT is applied in situations where usability testing is required, but the evaluator and users are separated in time and place. Consequently, participants can take part in the practical usability test at the time and place of their choice, which enables capturing realistic interactions with the target product. Separating observers and users in time and space makes it convenient to involve user groups in usability testing across organisational and geographical boundaries.

In the last decade, increasing attention has been paid to RAUT's capabilities. However, although the potential of RAUT as a formative usability testing method has been considered in the usability evaluation methods (UEM) literature, most previous studies have been comparative. Hence, the implications of applying RAUT to users in their natural remote environments have not been sufficiently investigated. The insights gained from research focussed on determining such implications could lead researchers and usability practitioners to better understand the capabilities and limitations of RAUT, as well as the expected level of validity of the data obtained from usability testing applied remotely to users in their natural environments.

The following sections of this chapter introduce the research, beginning with the background and context that have informed it. The following sections introduce the challenges and limitations of UEMs, concentrating on RAUT. The research motivation, the problem statement and the research questions are presented. The final section describes the organisation of the thesis.

1.2 Background

1.2.1 Challenges in UEMs Research

The majority of published accounts of usability evaluation were published two decades ago (Card et al., 1983; Nielsen and Molich, 1990), and comparative studies on UEMs were published even earlier (Nielsen and Molich, 1990). However, several challenges in the UEMs research have been reported in seminal papers by Gray and Salzman (1998), Hornbæk (2010) and Woolrych et al. (2011). These challenges can be summarised as follows: First, there is no agreement amongst practitioners regarding a uniform UEM or among researchers regarding a standard means for evaluating and comparing UEMs. Second, there is no understanding regarding the limitations of UEMs and when they are applicable for usage. Third, there is a lack of comprehension of how to conduct and compare UEM evaluations, which was pointed out by Gray and Salzman (1998) and agreed subsequently by Hornbæk (2010) and Woolrych et al. (2011). Hence, the results reported by these studies might be misleading (Gray and Salzman, 1998). Most UEM evaluation and comparison work has been limited by problems concerning validity, reliability and practical utility. Validity concerns limitations in the statistical tests and in the conclusions passed to practitioners and researchers, as well as in the measures used to compare methods. The reliability of the comparisons of UEMs is also questionable because of the evaluator effect (Hertzum and Jacobsen, 2001), which indicates that different evaluators find markedly different sets of usability problems[*] as a result of applying a particular UEM. Another issue is that most UEM evaluation and comparison work has focussed on discovering usability problems, neglecting the most important goal of UEMs, which is to evaluate design. This issue could lead to improper assessments of the practical utility of UEMs (Wixon, 2003). The fourth challenge

[*] We use the terms "usability problems" and "usability issues" interchangeably in this thesis. The term "usability problem" is used mainly as acknowledged in the literature or others, and the usability defects related to this research design will be referred to as "usability issues".

was raised 10 years after Gray and Salzman's (1998) paper, which concerned the implication of the focus on "win-lose" outcomes in the UEMs comparative studies literature (Hornbæk, 2010). Although much of the UEMs comparison work has been focussed on "win-lose" outcomes, in practice, usability practitioners appear to use a combination of methods rather than relying on the results of just one (Borgholm and Madsen 1999; Gulliksen et al. 2004). Assessments of UEMs to identify a "winner" do not provide helpful information for the practice of combining UEMs (Hornbæk, 2010). The choice of which UEM to use depends upon the kind of information the method is likely to offer.

The fifth challenge concerns overlooking contextual factors and their possible impacts (e.g., system fidelity, evaluator-developer gap, phase in development cycle, kind of system etc.). These contextual factors are all pertinent to understanding and evaluating the results of comparing UEMs (Hornbæk, 2010).

1.2.2 Limitations of Empirical Studies on RAUT

Although some efforts have been made to study RAUT methods, the knowledge of the contribution of the RAUT practice is inconclusive and incomplete. As described in the previous section, there is a lack of understanding of the capabilities and limitations of UEMs (Hartson et al., 2001), including studies that have evaluated RAUT or compared it with other UEMs. The first and the second challenges described in the previous section are common across almost all previous comparative studies that included RAUT. Additionally, these studies were conducted mainly to examine whether usability testing in laboratories could be replaced by remote settings (e.g., Bruun et al., 2009). This view of the comparison of methods (e.g., Andreasen et al., 2007) is based on the focus on "win-lose" outcomes, as discussed in the previous section (section 1.2.1).

In addition, the factors of validity, reliability and utility were considered in previous studies. The leading question in these studies was whether the compared UEMs yielded similar data. However, the findings of multiple studies differed greatly. For example, Tullis et al. (2002) found no difference in task completion between traditional lab usability testing and RAUT. Andreasen et al. (2007) also found no difference in task completion rate and task completion time between the two settings. Batra and Bishu (2007) found that remote usability testing did not differ from traditional usability testing; however, they did not describe the metrics they used in their comparison. In contrast, Andreasen et al. (2007) observed a significant

difference in the time spent on tasks between laboratory and remote settings, and Bruun et al. (2009) found that fewer usability issues were identified in the RAUT method compared with other methods.

The reason for these differences might be that the data were collected in different ways in lab and remote settings. Confounding the situation was that the results were referred negatively or positively to RAUT, but different innovations of RAUT were applied, such as user-reported critical incident (UCI) (e.g., Andreasen et al., 2007; Bruun et al., 2009) and web-based automated usability testing and questionnaire (e.g., Tullis et al., 2002; Batra and Bishu, 2007). The results were reported under the umbrella term of RAUT as the evaluation method used. Examples are comparisons of the completion time of RAUT when the UCI technique was used with traditional usability testing when the think-aloud protocol was used. Such comparisons are not valid, as the user-reported usability issues were collected differently in the two techniques. Similar to any usability evaluation method, all the aforementioned factors affect the validity of the data obtained with respect to RAUT.

According to Gray and Salzman (1998), comparative studies in the literature on UEMs are based on the perception that the compared or evaluated methods used are mainly formative UEMs. However, it appears that there was some confusion in the previous work regarding the involvement of the RAUT method(s). Because formative UEMs (e.g., laboratory-based usability testing) have a component with a summative component, they can also be used to gather quantitative usability data (e.g., task performance metrics such as time on tasks). Moreover, some previous UEMs comparison studies based comparisons, and their conclusions regarding which UEMs performed better, on quantitative data (e.g., Andreasen et al., 2007). For example, Andreasen et al. (2007) and Bruun et al. (2009) perceived the asynchronous usability evaluation as a formative UEM, but they were overly strict regarding the results of their data analyses, other than usability issues, such Andreasen et al.'s (2007) findings for time on task completion. The limitation of such studies was that quantitative data are not intended to provide the statistical significance usually required in summative evaluations (Hartson et al., 2001).

In addition, previous studies in the literature have been conducted from different perceptions and understandings of the term "remote" the test set-up, which led to differing results. Hence, the conclusions of comparative studies, especially with respect to quantitative measures, are not precise or valid. The insights gained from quantitative results might be valuable in the

usability engineering process in a local project. However, because they are not statistically significant, these results did not contribute (directly) to the science of usability (Hartson et al., 2001). That is, in formal comparison studies, analysing of quantitative and qualitative data should be treated with caution and awareness, depending on the objective of the research.

In general, there is a difference between conducting research on the effectiveness of a particular UEM in collecting data on the usability of a product or the usability of the interaction with a product, such as the practical application of some UEM, and comparing the data obtained to determine which are the best to use. The following practice was dominant in the UEM literature (Hartson et al., 2001):

> The inference about causality is very difficult to resolve in the case of UEM studies in which one is comparing one UEM against another that is potentially entirely different. The differences are far too many to tie up in a tidy representation by independent variables focusing us to compare apples and oranges. (Karat, 1998 cited in Hartson et al., 2001, p. 404)

Few researchers have described how they have collected their test data asynchronously from participants. As most remote studies have focussed on simulating laboratory usability testing in a remote environment, few attempts have been made to understand spatial and temporal differences between the evaluator's and participants' environments and their implications for the data obtained from usability testing. In most of these previous studies, contextual factors were overlooked. Andreasen et al. (2007) and Bruun et al. (2009) concluded that without information regarding distraction events, the interpretation of the data was difficult because "we [did] not know if the test subjects had any breaks during the test sessions, and therefore we [did] not know the exact time spent on the test" (Andreasen et al., 2007, p. 1410).

Bruun et al. (2009) stated the following:

> [O]ne of the difficulties in our study was that we did not observe the participants in remote conditions …. [T]he consequence is that we have missed information about the task-solving process. It also means that the task completion times have to be read with great caution. (Bruun et al., 2009, p. 1626–1627).

Some studies tried to exclude such factors as much as possible by considering them confounding variables. For example, in Tullis et al. (2002), the participants were provided with a pause button to stop the clock during task performance if they were interrupted or needed a break (Tullis et al., 2002). They also removed all data if a participant's task completion time was under five seconds or over 1,000 seconds because they considered such data to indicate either a lack of commitment (five seconds) or a possible interruption (1,000 seconds) (Tullis et al., 2002). Nevertheless, this perception of contextual factors in the users' remote environment resembled virtual laboratories even though the users' natural environments were not "transplanted replication[s] of laboratories" (Brewer and Crano, 2000, p. 14), rather than gathering data about the environment in which the UEM was actually applied (Hornbæk, 2010). Hence, most previous studies that have addressed asynchronous usability evaluation methods are considered UEM comparisons or/and evaluations. In other words, their results are based on comparisons of different methods according to the data obtained by each method. Therefore, the results of these studies should be considered with caution.

1.3 Research Motivation

RAUT needs to be revisited and reinvestigated for several reasons. Firstly, because of the potential of applying usability tests remotely (e.g., increased access to participants, reduced travel, lower expenses, automated testing etc.), the current body of the UEMs literature is insufficient. This is particularly true regarding the shortcomings of previous studies that have addressed RAUT methods, as described in the previous section.

Secondly, there is a need to address RAUT differently to gain insights into its capabilities. Researchers should focus on maximising the benefits of RAUT rather than simply comparing the different forms of RAUT methods to traditional lab usability testing or other usability evaluation methods, which has been the focus for almost two decades. Hornbæk (2010) argued that the best single method can only be identified if it is replicable. Moreover, it is difficult to replicate results across different systems and contexts because of resource constraints. Hornbæk (2010) further argued that focusing on comparisons and method innovations ignores the reality that usability evaluation methods are loose and incomplete collections of resources that successful practitioners configure, adapt and complement to match specific project circumstances. Considering the point raised by Hornbæk (2010), the research attention should be shifted to how we can maximise the benefits of target evaluation

methods and comprehend their shortcomings to maximise the amount of testing data provided by that method, rather than just compare it with other evaluation methods.

The third factor is that most UEMs comparative studies that included an asynchronous usability evaluation method considered it a formative usability evaluation method. Regardless of whether RAUT is effective in collecting data on usability issues, which is the main objective of the formative evaluation, it might be the only option for gaining insights into defects in user-product interactions in some projects, such as open source projects. Thus, RAUT needs more investigation.

The fourth factor is the concept of RAUT and its suitability for un-moderated automated testing techniques. Un-moderated automated testing is becoming increasingly important and used because of the additional advantages it provides in terms of the reduction in the time required to run studies with large numbers of participants and its capability of automated reporting and analysis. The capabilities of un-moderated automated testing make it ideal in applying summative evaluations, which are required to be applied repeatedly, need large numbers of participants to reach statistical significant levels, and must focus on the precise quantification of performance metrics of a finished product in comparison with a competitor's products or with different versions of the same product. Because of these traits, summative evaluation is ideal in remote automated delivery and administration. The automated un-moderated usability tools available in the market provide an objective and precise way to quantify performance metrics. Running summative evaluations in the traditional way (e.g., in a lab) can be time-consuming and expensive. In contrast, running summative evaluations through the use of RAUT in users' natural environments may mean that several layers of information may be lost, as no observer is present, which might affect the quality of the test data obtained. Clearly, there is a need for more research on RAUT.

1.4 Problem Statement

New communication technology has enabled the innovation and adoption of RAUT. Consequently, usability practitioners and researchers are able to reach users in any place and at any time. UEM research has been carried out mainly to compare the performance of RAUT in users' ordinary environments with other usability evaluation methods, such as the traditional lab usability testing method. Some results of these previous comparative studies

suggest that there are differences in the data collected on the performances of users who undertake traditional usability tests in labs and those who perform the tests remotely.

In addition, some comparative studies on UEMs involving RAUT have raised interesting points about the possibility of the existence of unknown contextual factors. However, these studies have not yielded insights into such contextual factors or the implications of their existence for the outcomes of usability testing. Those studies were merely focussed on trying to replicate the laboratory usability testing approach in ordinary environments and comparing the outcomes of RAUT with usability testing in the laboratory.

In the laboratory environment, we are fully aware of what happens during a usability testing session. However, when we apply usability testing with remote users, we have no indication of what might happen in their natural environment while they interact with the product during the usability test session.

Thus, to optimise the use of the RAUT method, we should not only rely on the fact that it enables users to be reached at any place and time but also be aware of what happens during the user's interaction with the product and the circumstances that surround the kind of user interaction in an uncontrolled environment. These circumstances may affect the quality of the data collected by RAUT and consequently the validity of the results. The awareness of such factors would enable the validation of the data collected from RAUT. Thus, RAUT needs to be investigated from a new perspective.

Based on the literature review and the above considerations, the main goal of this thesis is to gain insights into the implications of using usability testing with remote users[*]. Therefore, the research questions addressed in this thesis are as follows:

RQ1: What can we expect from the participants in remote usability testing when they are asked to report their own issues and outcomes?

RQ2: Does performance during usability testing in a (remote) natural environment differ from that of participants in a laboratory environment?

RQ3: What contextual factors are experienced by remote participants during their usability testing session?

[*] The term RAUT was used in the literature review and in the previous sections to refer to the literature, where it was generally called RAUT. However, as discussed in section 1.2.2, different methods were referred to as RAUT in the literature. Because this research focuses on the implications of remote application of the usability testing rather than investigating the method itself, from now on I use the general term "usability testing with remote users" for simplicity.

RQ4: How do the contextual factors influence the users' outcomes during usability testing?

RQ5: What is the effect or "the cost" of interrupting users' performance in usability testing on usability practice?

1.5 Overview of The Methodological Approach

To answer these research questions, this thesis will be based on an empirical approach, which will be described fully in Chapter 3. It is worth mentioning that this research does not compare UEMs. For example, it does not compare traditional laboratory usability testing and RAUT because of the problems with these kinds of comparisons, which were discussed in the previous sections. In this thesis, formal empirical summative online usability studies are used to answer the research questions using modern automated online tools. Empirical summative studies are used to compare performance metrics or design factors in a way that could add to the accumulated knowledge in the field of human computer interaction (HCI). Summative usability evaluations are suitable for un-moderated testing for many reasons. The nature of RAUT and the fact that it does not require an observer to be present makes it suitable for summative usability testing and online administration with remote users because it enables reaching remote users at any place and at any time.

Conducting a study on a usability testing method online should be formalised as an online study. In online studies, the internet is both a methodological tool used to administer a study and an object to be addressed (Orgad, 2009), which is referred to as internet research (Baym and Markham, 2009) or virtual research (Hine, 2006; Buchanan, 2004).

The advantages of online studies are that they enable accessing the usability study as long as there is an internet connection. From a practical perspective, administering usability testing online enables large number of users in globally distributed locations to be included in the sample. From an empirical perspective, in addition to enabling the recruitment of large numbers of participants, an online study can be run anywhere. Therefore, it can be used in empirical comparisons and experiments where identical or equivalent usability testing tasks are run to investigate a specific factor. The method itself is not the subject of the comparison. The method is fixed among different situations or experimental conditions, which are the investigated factor(s). This empirical perspective is adopted in this research.

The primary goal and challenge of my thesis is to investigate the implications of using remote usability testing with remote users. I therefore decided to address the above research

questions by conducting multiple experiments in the form of an empirical online summative usability study. I have adopted a two-stage approach in which the insights gained from the exploratory study applied in stage one serve as the basis for the design of the two empirical studies conducted in stage two. In each successive study, I have investigated or validated the identified reasons for the results in the exploratory study conducted in stage one. The two subsequent studies serve as explanatory and validation studies, respectively. The explanatory study provided explanations for the preliminary findings in the first study. The validation study both validated the second study's findings and provided more elaborate findings (Figure 1.1). The first exploratory study aimed to answer the first three research questions. The second explanatory study aimed to validate the answers provided by the first study, to answer the second, third and fourth research questions. The third research study aimed to validate the findings reported by the second study and to answer the fifth research question. Table 1.1 provides an overview of the research methodology. In Chapter 3, Table 1.1 will be elaborated on to provide additional context.

Table 1.1. Overview of Research Methodology

Study	**Methodological approach**	**Purpose / objective**	**Research questions to be answered**	**Research strategy**	**Dominant paradigm /perspective**	**Data type(s) collected**	**Dominant research approach**
Study 1	Empirical	Exploratory	RQ1 and RQ2	Comparative Observational	Postpositivist	Quantitative Qualitative	Quantitative
Study 2		Explanatory	RQ3, RQ4, and RQ5	Experimental Comparative Correlational	Postpositivist	Quantitative Qualitative	Quantitative
Study 3		Validation	RQ5	Experimental Comparative	Pragmatic Postpositivist + Constructive)	Quantitative Qualitative	Mixed mode / Triangulation (Quantitative + Qualitative)
In general, empirical research with pragmatic paradigm							

1.6 Structure of The Thesis

The remainder of this thesis is structured as follows.

Chapter 2: Literature Review

This chapter presents a background of usability and its evaluation methods, particularly usability testing, its approaches and its variants. Then the influential factors on usability testing are discussed. The early work on RAUT is then critically discussed. The chapter then presents background information about distractions and discusses how they are addressed in the literature.

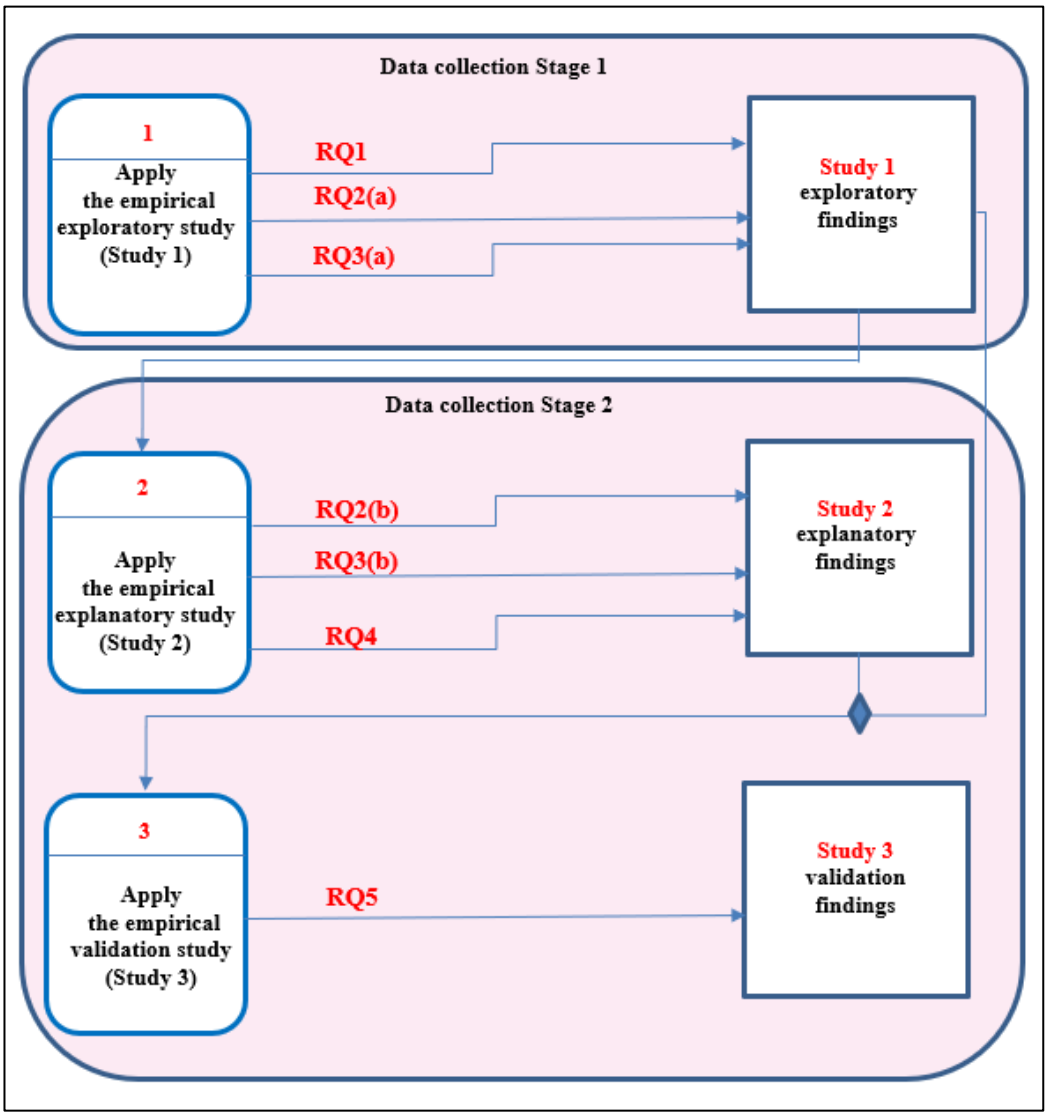


Figure 1.1. Overview of the methodical approach

Chapter 3: Methodology

This chapter seeks to justify the choice of the methodology used in this study through a general discussion of the underlying research paradigm and a description of the main research method and its design. The chapter then discusses the factors considered during the experimental design phase, the methodological techniques used in the collection of the empirical data, and the strategies used to analyse the data. Lastly, it describes the research design based on the formulated theoretical framework and rationale for the methodology.

Chapter 4: Empirical Exploratory Study

This chapter presents the empirical exploratory study, which is aimed at exploring the functionality of usability studies in administering the test, its tasks, instructions and questions in different experimental settings. The chapter presents the data provided by the participants through the online administrated usability study during testing sessions in different testing environments. The chapter presents the preliminary findings on the usability outcomes in different testing environments. The limitations and implications for further studies are discussed. This study is intended to answer the first research question and address the potential of the second and third research questions (Table 1.2).

Chapter 5: Empirical Explanatory Study

This chapter presents the empirical explanatory comparative study, which is aimed at investigating the usability testing outcomes of the participants' performance and their subjective reports in laboratory and natural environments. It also investigates the contextual factors experienced and reported by the participants in the natural environment and whether there is any relationship between the usability testing outcomes and the contextual factors reported. This study is intended to answer the second, third and fourth research questions (Table 1.2).

Chapter 6: Experimental Validation Study

This chapter presents the final empirical study, which aimed to validate the findings of the exploratory and explanatory studies. In particular, this chapter reports an experiment that was designed and conducted to investigate the cost, that is, "the influence" of interrupted task performance in usability testing. This study is intended to answer the fifth research question (Table 1.2).

Chapter 7: Discussion

This chapter provides an evaluation and discussion of the main findings of this research.

Chapter 8: Conclusion

This chapter concludes the thesis by summarising the concepts developed and the contributions of the research. In addition, it provides suggestions for extending the research in the future.

Table 1.2. Contribution Chapters, Their Associated Empirical Studies and the Research Questions Addressed

Chapter	**Study sequence**	**Purpose**	**Research questions addressed**
Chapter 4	Study 1	Exploratory	RQ1, RQ2(a), and RQ3(a)
Chapter 5	Study 2	Explanatory	RQ2(b), RQ3(b), and RQ4
Chapter 6	Study 3	Validation	RQ5

Chapter 2: Background and Literature Review

2.1 Overview

The research problem and research questions were introduced in Chapter 1. This chapter presents background information about usability and the methods used to evaluate it. This is followed by a description of usability testing, its approaches and its variants. The chapter then discusses factors that have been found influence usability testing. The literature on RAUT and previous studies that attempted to investigate the influence of the environment on usability testing outcomes are reviewed and discussed. The chapter then presents background information about distraction and discusses how it is addressed in the empirical literature.

2.2 Background

2.2.1 Usability

"Usability" is a construct conceived by the HCI community to denote a desired quality of interactive systems and products. Three international standards have defined usability (Table 2.1). The World Wide Web has become a prevailing and dominant interface. This is a result of the exponential growth in the number and the size of e-business and e-governments sites, for instance, which answered the need for applying the basic usability principles to the web environment. Therefore, usability researchers have developed standards, guidelines, tools, and technologies for web use (Tung et al., 2009).

The most applicable definition of usability in the context of Web usability is that of ISO9241-11 which refers to "the extent to which web sites can be used by specified users to achieve specified goals to visit with effectiveness, efficiency, and satisfaction in a specified context of website use" (ISO9241-11, 1998, p.170). The usability and design of Web sites has received attention in HCI literature as well as in Web-specific usability research.

Usability has typically taken an engineering approach in an attempt to identify a set of principles and common practices that will ensure usability is an outcome of system design (Nielsen 1993, Pearrow 2000, Zhang et al., 1998).

Table 2.1: Definitions of Usability According to Different Standards

Standard	Usability definition
(IEEE, 1990, p.80)	"The ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component."
(ISO9241-11, 1998, p 170)	"The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."
(ISO/IEC 91260-1, 2000)	"The capability of the software product to be understood, learned, used, and attractive to the user, when used under specified conditions."

Nielsen's definition of usability/usability model consists of five attributes: learnability, efficiency, memorability, errors and satisfaction. According to Nielsen, learnability indicates how easy the system is to learn. Learnability can be measured by counting the number of correct steps when performing a particular task after the first time. Efficiency concerns the ability of the user to complete the task within an acceptable amount of time and it could be measured by calculating the time consumed to complete a task. Memorability means that the system functions should be easy to remember, so that a casual user can return to the system without relearning how to use it. It could be measured by counting the number of steps remembered and performed by the user in the second usage. Usability implies that the evaluated system should be having a low error rate which could be measured by counting the number of errors made by the user while performing a specific task. Satisfaction means that the system should be pleasant for the user, which will be reflected in user satisfaction. Satisfaction can be assessed by subjective, qualitative inquiry into whether the user was happy with the system (Nielsen, 1993). Nielsen's attributes have been applied in many different studies including website usability studies (Downing & Liu, 2011).

The ISO9241-11, (1998) definition for usability is more generic and includes only three primary factors which are: effectiveness, efficiency and satisfaction. Effectiveness characterises the completeness and accuracy in users' performance (e.g., information gathering, purchasing) while surfing a website (Tripathi et al., 2010). It is directly related to the right functionality so that users can do what they need or want to do while visiting a website.

The second factor is efficiency, which represents the resources expended in relation to achieving goals while visiting a website. The users perceive efficiency when they can achieve goals with a quick visit without putting in much cognitive effort. The last factor is satisfaction which is defined as the comfort and acceptability of a website to its users.

Website usability is considered a multidimensional construct that encompasses effectiveness, efficiency and satisfaction due to website design.

Both these definitions, of Nielsen and ISO, have been considered a base for achieving the usability of a website (Downing, & Liu, 2011). Yet, other standards and models have also defined similar or different attributes. Refer to the following table, Table 2.2.

Table 2.2: Usability Attributes According to Different Standards/Models

Standard	**Nielsen (1993)**	**Preece et al., (1994)**	**ISO 9241-11 (1998)**	**Quesenbery (2001)**	**Shneiderman and Plaisant (2005)**
Attribute	Learnability	Learnability	Effectiveness	Easy to learn	Time to learn
	Efficiency	Throughput	Efficiency	Efficient	Performance
	Satisfaction	Attitude	Satisfaction	Effective error tolerant	Satisfaction
	Errors			Engaging	Errors
	Memorability				Retention

Information about the usability of a system is typically investigated in order to assess it—this practice is called usability evaluation. According to Fitzpatrick (1998, p.2), a usability evaluation method is a 'systematic procedure for recording data relating to end-user interaction with a software product or system'.

The data gathered from the evaluation process is analysed and assessed to determine the usability level. According to Dix et al. (2004) there are three general goals of the assessment: evaluate users' experience of the interaction with the system, identify the system's problems during a specific task and evaluate the system's functionality (Dix et al., 2004).

2.2.2 UEMs

There are different perspectives in the literature to classify usability evaluation methods. One perspective to classify the UEMs is based on the evaluation objective, to be either formative or summative. In the context of usability, the objective of the formative usability evaluation is to find the usability problems so that an interaction design can be fixed during development to improve the system design. While for the summative evaluation, the objective is to assess or compare the level of usability achieved and it takes place after development to assess the design (absolute or comparative) (Harston et al., 2001).

Another perspective of usability evaluation is based on how the evaluation was done, so it can be analytical or empirical. Analytical evaluation is based on analysis of the

characteristics of the design through examination of a design presentation, prototype, or implementation. Empirical evaluation is based on observation of performance of the design in use (Hix and Hartson, 1993).

According to Dix et al. (2004), evaluation can be categorised according to the location, for example, the normal, working environment or the laboratory. Lewis and Rieman (1994) divided the approach to evaluation according to whether the system was assessed with or without the user. Table 2.3 below summarises different categorisations of usability evaluation methods.

Table 2.3: Categorisations of Usability Evaluation Methods

	Categories
Faulkner (2000)	Formative
	Summative
Hix & Hartson, (1993)	Analytical
	Experimental
Dix et al. (2004)	Laboratory
	Natural Environment
Lewis & Rieman (1994)	User involved
	Without user

In practice, one or more evaluation method should be applied in the usability evaluation stage of the system development cycle (SDLC)—depending on the assessment aim—in order to discover usability problems and/or to measure users' performance in reaching the goals of a certain task. Several authors have identified a number of different evaluation methods (Preece et al., 1994; Shneiderman and Plaisant, 2005; Dix et al., 2004), some of which require the involvement of users, and others that require the involvement of experts in the field (Anandhan et al., 2006). The choice of usability evaluation to be used is typically based on the objective of the evaluation, the type of the system to be evaluated, the cost, time constraints, and appropriateness.

Table 2.4 presents an overview of the various usability evaluation methods, followed by a discussion of each method. Since this thesis concerns the implication of applying usability testing with remote users, and usability testing will be used as the experimental design method, it will be particularly detailed in the following section.

Table 2.4: Overview of Usability Evaluation Methods

Usability Method Type	Evaluator	Example of techniques	Evaluators' role
Model based	Expert	GOMS Parallel design	Use model to extract usability measures.
Inspection	Expert	Cognitive walkthrough Card Sorts Heuristic evaluation	Review the examined user interface to identify the problems.
Testing	User	Thinking aloud Observation Co-discovery Remote/Field testing	Observe users using the system. Analyse the collected data to explore users' performance, usability issues, and/or users' usability assessment.
Inquiry	User	Interview Focus groups Questionnaire/Survey	Asked the users to get insights to define the problems and/or assessment for usability level.

2.2.3 Usability Testing

Usability testing is a user-based testing process that involves representative users who attempt to complete representative tasks (Lazar et al., 2010). According to Preece et al., (1994), it is an adapted form of experiment designed to test the usability of a system (Preece et al., 1994). Usability testing can take place very early or very late in development. Ideally, usability testing is conducted during all stages of development, but it is not always possible. Usability testing is widely regarded as the most fundamental and important method for identifying problems in user-product interactions (Nielsen, 1993).

2.2.3.1 Usability Testing Approaches

In conducting usability tests, designers must use usability metrics to specify what they intend to measure. Metrics are variables that are specified according to the scope and goals of the project. Exploratory usability testing, which typically takes place early in development, is also known as formative testing. It tends to be informal, and there is more communication between the test moderator and the participants. Exploratory usability testing usually uses inexpensive low-fidelity prototypes in small user groups of designers and users in an interactive and comfortable atmosphere (Rubin and Chisnell, 2008). Such usability testing concerns user satisfaction, as the focus is on how the user perceives the interface rather than how well the user completes the tasks (Rubin and Chisnell, 2008).

Usability metrics are quantitative with a refined or functional prototype that uses sophisticated testing equipment, such as high-fidelity. This kind of usability testing is

summative testing, which concerns effectiveness, efficiency or/and subjective satisfaction, as the focus is on evaluating the effectiveness of the interface design (Dumas and Fox, 2008). Data on these usability issues are typically collected by asking users to complete various tasks using the target system. Effectiveness metrics can be measured through successful completion rates.

Whether usability testing is formative or summative affects how formal or informal the usability test is. At one end of the chain is the formal approach to usability testing, which parallels experimental design. Formal usability testing requires specific research questions, research design, and multiple design interfaces. In addition, if this usability testing involves inferential statistics, it may require a control group and a large number of subjects, which represents the experimental design of a user study. The difference between experimental design and practical usability testing is that the former is conducted to determine statistically significant differences between groups, whereas usability testing is conducted to find ways to improve specific interfaces (Lazar et al., 2010).

2.2.3.2 Usability Testing Variants

The review of the literature on the types of usability testing revealed that there are two views of usability testing techniques. The first view represents the traditional view of usability testing techniques (e.g., Lewis, 2006) which is based on the methodological and technical aspects of the technique used to collect measurement data from users. The second, more recent view of usability testing (e.g., Lazar, 2010) is based on the location of the test and how it is set up.

- **Technical aspects**

Usability testing can be applied using the following techniques: the think-aloud protocol (TAP), observation, co-discovery or remote usability testing. These techniques are either synchronous or asynchronous. TAP has been defined as a type of empirical research that asks users to perform a task and verbalise their thoughts during the task (Jaäskeläinen, 2001). According to Ericsson and Simon (1998), TAP is a valid method for analysing cognitive processes, as it accesses the users' issues and thoughts arising in their short-term memory during testing. This method is considered advantageous because it elicits data from short-term memory, which is unaffected by users' perceptions (Ericsson and Simon, 1998), and it

can be used effectively with minimum training (Nielsen, 1993). However, users' utterances are often incoherent (Ericsson and Simon, 1998), and they might not be able to express their thoughts freely (Van den Haak and de Jong, 2005), which might be related to the cognitive load induced by problems in speaking in some study participants (e.g., Branch, 2000). Although TAP is typically conducted in a laboratory, the recent availability of screen sharing and recording technology has meant that it can be applied remotely with users in their natural environment.

Using observation tools, data are collected from actual users while they interact with a system. The investigator monitors users while they perform the required task and makes notes about their activities. The method is useful for obtaining qualitative data, and it can be combined with other inquiry methods to achieve even more useful results. It is considered simple compared with other usability testing techniques, as it does not require additional software or tools. This method can be applied either in the laboratory or in a working environment (Preece et al., 1994).

In co-discovery learning, two users are observed while they work together to perform a specific task. This technique is considered more natural than TAP because the two users share thoughts while performing the task, which is considered a natural discussion (Zaphiris and Kurniawan, 2006). According to Nielsen (1993), it is preferable to pair two subjects who know each other well to ensure that they feel comfortable discussing issues; however, this requirement cannot always be achieved.

The improvements in networking and communication technologies have given rise to the application of remote communications techniques with the usability testing method. The usability testing applied with these means of communication has been termed "remote usability testing". It was defined as evaluations of users who are in different locations (Ivory and Hearst, 2001). Remote usability testing techniques are generally classified as either synchronous or asynchronous.

In the synchronous technique, users and evaluators are separated spatially. In the asynchronous technique, users and evaluators are separated in both space and time (Andreasen et al., 2007). Remote usability testing provides a vehicle for easily soliciting feedback from users in remote areas. Remote usability testing can provide both quantitative and qualitative data. Synchronous techniques (also known as moderated) are usually used in remote usability testing in qualitative studies to validate suspected usability issues. Recent

synchronous techniques allow for observing a subject's screen and verbal "think-aloud" commentary (screen recording video) and enable capturing webcam views of the subjects (video-in-video [ViV]). However, these tools are costly. The asynchronous technique (also known as unmoderated) usually includes the use of a specially adapted online survey, which allows quantitative user-testing studies, which enables the generation of large sample sizes. According to Albert et al. (2009), attitudinal data and, to some extent, behavioural data can be collected using this technique, such as through an online usability study. This technique can provide an opportunity to segment feedback according to demographic, attitudinal and behavioural types. These tests, which are carried out in the user's own environment rather than a laboratory, help to further simulate real-life scenario testing although they have been recognised as being harder to control (Lazar et al., 2010).

- **Usability test location**

Usability testing can be applied anywhere, such as in a fixed laboratory, a workplace, a user's home, over the phone or over the Web. The decision of where to conduct the usability test should be formed based on locations that are available, the participants' location, the purpose of the project or test, and the type of data to be collected. Therefore, no location is superior to any other location (Lazar et al., 2010).

Traditionally, usability testing takes place in a laboratory. The laboratory setting can range from the most formal setting, which is a two-room set-up, to one evaluation room. In the two-room set-up, a user sits in one room and performs tasks; his/her performance is recorded using a microphone and camera in addition to his/her computer screen. The moderator and possibly other stakeholders sit in another room and watch the user's performance via computer screens and the recording equipment. The moderator can directly observe what the user is doing through a one-way mirror, but the user cannot see the moderators' room. In the one evaluation room setting, the moderator sits with the user, who is positioned to minimise distractions but to maximise view (Lazar, 2006; Murphy et al., 2007).

Usability testing can take place in the users' workplace or home. This approach provides simple user recruitment, as they do not have to travel to a usability laboratory or central location. It also helps users with impairments for whom transportation is challenging. In this set-up, the user is exposed to everyday distractions, noise and attention limitations. However, users may feel comfortable because they perform the test in their normal environment. The

test can be set up in different forms. In the most challenging form, the test moderator needs to visit each user's workplace or home. In the easiest form, the usability practitioner/test moderator administers the usability test online over the Web (i.e., website) and allows the users to perform the test at a time and place of their choice (Lazar et al., 2010).

When the test takes place in a user's workplace or home, the test moderator must decide whether he/she wants to install the software or interface on the user's computer or bring his/her laptop with software or interface installed on it. The former is a more natural test, yet more technical problems might occur. Whether to apply the observation technique or data recording is another decision that must be made by the test moderator. There are different approaches, all of which have both benefits and drawbacks: direct observation, which might place influential factors on the user's performance; data logging (the user's keystrokes that are recorded); and audio and/or screen recording. Another option is to use a portable usability laboratory, which includes the same equipment as in a fixed usability laboratory. However, this solution is likely to be costly, and it is not guaranteed to avoid all technical problems (Lazar et al., 2010).

The easiest form of usability testing is one that enables representative users to participate in the usability test in their natural environment. In this form, the moderator finds that it is not feasible to do usability testing in a centralised location at a usability lab or to travel to a user's workplace or home because of logistical limitations that hinder the ability to apply face-to-face usability testing. Examples are situations where the representative user population is not within easy travelling distance of the usability evaluators or moderators; the test is meant to be done with individuals with disabilities for whom transportation might be a problem (Petrie et al., 2006); it is not possible for the evaluators to visit all the countries where the interface needs to be evaluated (Dray and Siegal, 2004). In such situations, video, audio and network connections allow testing evaluators to monitor users, including streaming the output from the user's screen (Hartson et al., 1996). This type of testing is called "remote usability testing", which was discussed earlier. However, excellent connections are necessary when testing is conducted through video conferencing on a private network or through a broadband connection to the Internet. In addition, observing non-verbal and interpersonal cues is challenging (Dray and Siegel, 2004). Overall, remote usability testing is regarded as more appropriate in summative testing that involves quantitative

metrics than for formative testing that involves qualitative observations (Dray and Siegel, 2004).

This thesis focuses on usability testing with representative remote users in their natural environments where interactions are recorded and logged using online means. It has been suggested that the outcomes and/or the data of usability testing in such situations might be influenced by certain factors (Dray and Siegel, 2004), which will be discussed in the following section.

2.2.3.3 Usability Testing and Influential Factors

Several researchers have discussed factors that influence usability testing outcomes. Some have discussed user numbers, their characteristics and how they influence usability testing outcomes. For example, the influence of the number of users on usability testing outcomes represented in usability issues revealed (e.g., Nielsen, 2000; Lindgaard and Chattratichart, 2007) the influence of user experience and familiarity with the system, as more users might be needed if the target website were new to the users (Lindgaard and Chattratichart, 2007). Another factor discussed in the literature has been task design, such as the influence of a detailed task description on the testing results (Sears and Hess, 1999) and the influence of task design selection on the evaluator's role in terms of problem detection and therefore usability problems (e.g., Hertzum and Jacobsen, 2001).

Additionally, the prototype fidelity of the target system has been discussed thoroughly in the HCI literature. The description of the HCI community's view of prototype fidelity was detailed by Rudd et al. (1996). Two design fidelity categories are generally used in categorizing prototypes: low-fidelity and high-fidelity. Some researchers have discussed the influence of prototype fidelity on the outcomes of usability testing. An example is the influence of the type of prototype fidelity on the type and number of usability issues (Nielsen, 1990; Virzi et al., 1996).

With regard to the factor of the testing environment, except studies by Andrzejczak and Liu (2010) and Greifeneder (2011), the influence of the testing environment on usability testing outcomes has rarely been discussed in the HCI literature. The environment factor was usually considered a methodological factor in comparative studies that investigated which usability evaluation method would work better: users perform more efficiently and effectively; the evaluation would reveal more usability issues. However, conflicting findings

were often acknowledged in these studies. For example, the findings in Andrzejczak and Liu (2010) conflicted with those in Greifeneder (2011).

It is clear that multiple factors can influence usability testing outcomes. However, in trying to characterise or modulate usability testing, most previous research focussed on technical system fidelity (e.g. prototype fidelity) but overlooked other contextual factors, testing environments and user characteristics.

For example, Nilsson and Siponen's (2005) model characterises three aspects of fidelity: implemented automaticity (i.e., the degree to which a user can operate a prototype without the test facilitator's assistance); perceived automaticity (i.e., the subjective assessment of automaticity level); and precision (i.e., the level of detail at which a prototype is modelled).

Virzi et al.'s (1996) model is based on prototype fidelity but with a somewhat broader understanding. It encompasses four dimensions: degree of functionality (i.e., to which details in a function are modelled); similarity of interactions (i.e., the level of mapping HCI, communication, and the type of displays and controls); aesthetic refinement (i.e., the product modelling regarding colours and shape), feature breadth (i.e., feature quantity in a modelled prototype).

Elliot et al.'s model (2004) is based on prototype fidelity, but it provides a much broader view of fidelity, which includes aspects of fidelity that are not limited to prototype design. The model includes other aspects, such as task characteristics (e.g., distributed team tasks) and operational requirements (e.g., mission goals).

The review of these models further suggests that none explicitly considers the wider testing environment in which HCI takes place. The usability testing literature has acknowledged the importance of the wider usage context (Nielsen, 1993; Snyder, 2003), yet the focus has been mainly on the system itself. In addition, with respect to user characteristics and testing environment factors, relatively little guidance has been given to designers regarding the fidelity level to be used.

Trivedi and Khanum (2012) suggested that similar to product characteristics, a usability evaluation model should encompass context characteristics (i.e., the users, tasks, and environment) in determining usability. According to Bevan and Macleod (1994), changing any applicable characteristic of the usage context may alter product usability. The usage

context can include cultural context (Nivala and Sarjakoski, 2003), organisational context, technological context and social context (Maguire, 2001).

In the four-factor framework of contextual fidelity (4FFCF) model (Sauer et al., 2010), a wider view of fidelity is proposed, which is not limited to the prototype, as it considers the fidelity of the entire context of the usability test. In the 4FFCF model, context fidelity is characterised by four main factors: *system prototype, testing environment, user characteristics* and *task scenarios.* Each factor is further defined in sub-factors (see Figure 2.1). The 4FFCF model extends the previous models and addresses pertinent issues discussed in the usability literature (e.g., user experiences) and issues that play a role in ergonomics beyond usability (e.g., social and physical environment).

In the usability testing context, the 4FFCF model is surrounded and influenced by multiple factors that might affect its outcomes (see Figure 2.1). Framework factors need to be empirically tested to investigate their influence on the outcomes of usability testing, which should be carried out after estimating the factors that influence usability testing outcomes. These factors are important because of the high possibility of their influence on user behaviour during usability testing and therefore its outcomes. Additionally, these outcomes may differ in settings where the user may exhibit varying behaviours. That is, in evaluating the usability of any system, the behaviour of the user must be considered. Consequently, contextual factors may violate the reliability and validity of the usability test. In psychological testing, reliability and validity are important principles to maintain, which also apply to usability testing and the participants involved in the test. In addition, the objectivity of the testing procedure is important, as well as how the test outcomes are recorded and how the results are interpreted.

According to the 4FFCF model, environmental factors consist of the social testing environment and the physical testing environment. The social testing environment refers to the presence of other people while the usability test is conducted (e.g., evaluators or facilitators) and its potential influence on test outcomes, following the social facilitation theory (Zajonc, 1965), according to which the presence of observers may influence appliance operation in usability testing.

The physical testing environment refers to several characteristics, such as the distractions, noise levels and location of the environment in which users participate in the test.

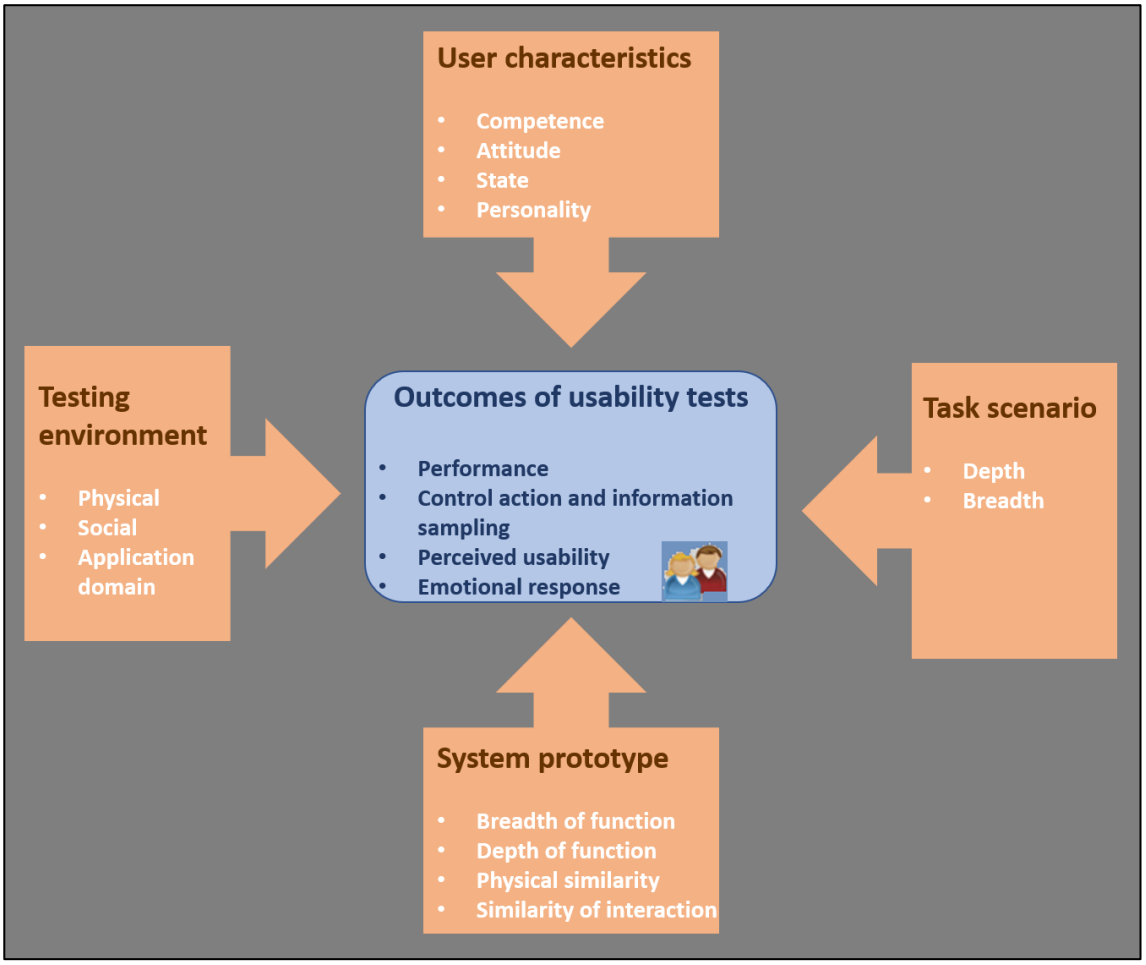


Figure 2.1. The Four-Factor Framework of Contextual Fidelity (4FFCF) (Source: adapted from Sauer Et Al., 2010, P. 132)

The environment where the system is typically used is called the "natural environment" (Trivedi and Khanum, 2012). The physical testing environment may influence user behaviour, which was shown in previous work on physical stressors (McCoy and Evans, 2005).

The behaviour setting theory proposes that precisely identifiable environment units, mainly physical and social elements, are integrated into one unit, and they highly influence human behaviour (Scott, 2005). Considering environmental influences on usability testing outcomes is important because of the inconsistencies found in the data collected by usability tests in the literature (e.g., Kessner et al., 2001; Lewis, 2006; Molich et al., 2004).

The validity, reliability and objectivity of usability tests in these previous studies are questionable because their outcomes may vary noticeably across tests, observers and methods. Accordingly, it is highly likely that the conflicting outcomes of usability tests are to a certain extent caused by uncontrolled and not well-understood features of usability tests

(Sauer el al., 2010). In light of the 4FFCF model and RAUT, we can easily consider the testing environment to be the most prevalent factor that might influence their outcomes. The reason is that testing environments can vary widely. For example, a users' natural environment is prone to distractions. The influence of social distractions was studied by Sauer and Sonderregger (2009), who found empirical evidence that the presence of observers in conventional laboratory usability tests may have negative effects on physiological parameters and on some aspects of performance. This thesis seeks to contribute to filling the gap in the knowledge regarding the physical testing environment influence on RAUT outcomes. Indications of the potential influence of physical environmental factors on RAUT and the lack of attention to them in the HCI literature are discussed in the next section.

2.3 Literature Review

In the previous subsection, we discussed how usability testing could be influenced by different factors. However, these factors have rarely been investigated or studied in the literature, especially the influence of environmental factors on usability testing outcomes. It could be supposed that social environmental factors could influence usability testing in a laboratory because a moderator or observer is present (i.e., being observed in an unfamiliar location) (Sauer and Sonderegger, 2009). However, the physical environment might be highly influential when the participants perform the usability test in their natural environment, which is most likely to be uncontrolled and open to distractions and noise, as stated previously.

In the following subsections, previous studies on RAUT will be reviewed first. The purpose of this critical review is to highlight the issues overlooked in these studies, particularly regarding validity. Then the subsequent section will focus on the few studies that sought to investigate the influence of the test environments on testing outcomes, as well as the contribution of this research to filling the knowledge gap. Because distraction is acknowledged in the literature as the predominantly influential contextual factor in task performance, previous studies in the literature on distraction have also been reviewed.

2.3.1 Earlier Investigations of RAUT

In addition to the earlier work on exploratory empirical applications of RAUT methods (Table 2.5), most of the work on RAUT has been in the form of comparative empirical

studies. These comparative studies compared RAUT methods with conventional laboratory usability testing and other evaluation methods. Examples are Tullis et al. (2002), West and Lehman (2006), Andreasen et al. (2007), Batra and Bishu (2007), Bruun et al. (2009), and Kelly and Gyllstrom (2011). The main objective of these studies was to examine whether the compared method could replace laboratory usability testing or to suggest the best method to use. However, the results of these comparative studies were inconsistent, such as the findings regarding task completion time. For example, Tullis et al. (2002), Andreasen et al., (2007) and Batra and Bishu (2007) found no difference in task completion time between the outcomes of usability testing in the lab and remote settings. However, Bruun et al.(2009) remarked on a considerable difference in task completion time between laboratory and remote settings. With respect to the number of usability issues, Andreasen et al. (2007) and Bruun et al. (2009) found fewer usability issues in remote applications of RAUT than in other methods.

The reasons for these differences in the results of previous studies might be two main issues: validity and environmental factors. Regarding the first issue, in UEM comparative studies, the collected data were often recorded, observed and quantified differently among the compared methods. For example, all the remote usability testing outcomes were referred to as "asynchronous usability evaluations", and different types of usability evaluation techniques were used to record asynchronous data collected from users in their remote natural environments (Table 2.5).

For example, some studies used the UCI technique to collect data on usability issues (e.g., Andreasen et al., 2007). Some used auto-logging (e.g., Bruun et al., 2009) to collect data on other performance metrics (e.g., task time and successful completions) and others used Web-based automated usability testing (e.g., Tullis et al., 2002) to collect data on task time and successful completion along with questionnaires to collect data on task time and successful completions (Tullis et al., 2002; West and Lehman, 2006; Batra and Bishu, 2007). For example, in the UCI technique, the time at which the users report the incident is included in the time per task measurement because they typically report the incident directly as it happens. In online-survey based testing, the users give feedback on the usability of the website and the issues they encountered after the task at the end of the test.

Table 2.5. Categorisation of Earlier Investigations Of RAUT: (A) Empirical Application of The Method, and (B) Empirical Comparison of The Method*

		Methods Used			
(A)Empirical Application using		**Questionnaire**	**UCI**	**Auto-logging**	**Unstructured problems reporting**
		(1) Hartson and Castillo (1998) (2) Ericsson and Simon (1998) (3) Winckler et al. (2000) (4) Tullis et al. (2002) (5) West and Lehman (2006) (6) Andreasen et al. (2007) (7) Batra and Bishu (2007) (8) Symonds (2011)	(1) Hartson et al. (1996) (2) Castillo et al. (1998) (3) Hartson and Castillo (1998) (4) Andreasen et al. (2007) (5) Bruun et al. (2009)	(1) Millen (1999) (2) Scholtz (1999) (3) Winckler et al. (2000) (4) Bruun et al. (2009)	(1) Ericsson and Simon (1998) (2) Äijö and Mantere (2001)
(B) Empirical comparison with:	**ARLT**	**(1), (6)**	**(5)**	**(4)**	**NA**
	TSLT	**(1), (6)**	**(5)**	**(4)**	**(2)**
	ARI	**(6)**	**(1), (2), (3), (4), (5)**	**NA**	**(2)**
	TSI	**NA**	**(1)**	**NA**	**NA**

* The numbers in parentheses are identifiers of the work cited in the same column of the first part of the table (A), and mean that the empirical results of that work are compared with (B), the results of the same work applying Asynchronous Remote usability Testing (ARLT), Traditional Synchronous Lab Testing (TSLT), Asynchronous Remote Inspection (ARI) or Traditional Synchronous Inspection (TSI).

In TAP, users are encouraged to verbalise their thoughts during task performance, which might increase the time to complete tasks. Moreover, different perceptions and understandings of the term "remote" and the set-up of the test may cause different results to be obtained. Sample size is also an issue in the reviewed studies. For example, in Brush et al. (2004), the sample size was eight participants in the laboratory and twelve participants in the remote setting. In Thompson et al. (2004), the sample size was five participants in both settings, and West and Lehman (2006) reported 17 participants in the laboratory setting and 13 in the remote setting. Bruun et al. (2009) recruited 10 participants for each setting (i.e., laboratory, UCI, diary and forum). Andreasen et al. (2007) recruited six participants for each setting (i.e., lab, remote synchronous usability testing, RAUT and remote asynchronous expert testing). Exceptions were Kelly and Gyllstrom (2011), who used a sample size of 30 participants in a laboratory setting and 39 participants in a remote setting. However, even with the reasonable number of participants recruited in the last three studies, the statistical validity of their conclusions is questionable because no information was reported on how the heterogeneity of participants was ensured. All the aforementioned factors can affect the validity of a comparison.

The second issue concerns the possibility of the presence of influential contextual factors on usability testing outcomes, which in RAUT is mainly the presence of distractions. Some studies demonstrated the awareness of physical environmental factors that could cause variability in comparison outcomes. For example, in Tullis et al. (2002), the participants were provided with a pause button to stop the clock during task performance if they were interrupted or needed a break (Tullis et al., 2002). They also removed all data if a participant's task completion time was under five seconds or over 1,000 seconds because they considered such data to indicate either a lack of commitment (five seconds) or a possible interruption (1,000 seconds) (Tullis et al., 2002).

In addition, in Kelly and Gyllstrom (2011), the remote setting was a virtual laboratory and distraction was considered an extraneous variable and thus excluded from the analysis. The participants were informed that "they should complete the study in one uninterrupted session, close all other applications on their computers and not multi-task" and that they should refrain from:

> answer[ing] their cell phones and/or reading/sending text messages.... [T]he system would automatically log them off after a 10-minute period of inactivity, and they would not be able to resume the study later (Kelly and Gyllstrom, 2011, p. 1534).

Andreasen et al. (2007) and Bruun et al. (2009) affirmed that without information regarding distraction events, the interpretation of data was difficult because "we do not know if the test subjects had any breaks during the test sessions, and therefore we do not know the exact time spent on the test" (Andreasen et al., 2007, p. 1410). Bruun et al. (2009) stated the following:

> [O]ne of the difficulties in our study was that we did not observe the participants in remote conditions.... [T]he consequence is that we have missed information about the task-solving process. It also means that the task completion times have to be read with great caution (Bruun et al., 2009, pp. 1626–1627)

2.3.2 Earlier Investigations of The Influence of Testing Environments on Usability Testing Outcomes

A few previous studies investigated the influence of different testing environments on usability testing outcomes. For example, Andrzejczak and Liu (2010) investigated the effect of test location (lab vs. remote) on usability testing performance, participant stress level, and subjective testing experience. They adopted UCI reports in the remote setting, and the test was applied synchronously.

Khanum and Trivedi (2013) investigated the effects of the testing environment on usability testing outcomes using TAP with children in an unfamiliar lab room and a familiar computer lab in a field setting, an approach similar to the local remote testing described by Hartson et al. (1996). Both studies observed the high possibility of distractions in the remote field environment. Andrzejczak and Liu (2010, p. 1265) stated, "Distractions and stressors may be present and not controlled in the remote laboratory setting such as disruptive students, fire drills, and other distractions present in a high-traffic environment". Khanum and Trivedi (2013, p. 2052) stated, "In the field test, there were interruptions as no restrictions were imposed on the people to move in the field, but these did not affect the performance much". However, neither study attempted to gather data on these distractions in order to relate differences found, if any, to them.

In contrast, Greifeneder's (2011) study was conducted in both settings, lab and remote, and was applied and administered online. Her study gathered data about distractions during the natural environment session and attempted to determine whether there was a relationship between the distractions reported and the differences found. However, it could not be concluded whether the few differences found were caused by the contextual factors reported by the participants in the remote setting.

This thesis aims to fill the gap in the knowledge about all these factors and issues by drawing inferences and addressing contextual factors that might influence the outcomes of usability testing applied to remote users, while avoiding or at least mitigating the validity issues in UEM comparisons. The conclusion to this thesis would provide insights into the implications of applying usability testing to remote users in their natural environment for the practice of usability testing. How could such insights be attainable? The answer to this question would demonstrate the novelty of this research, which will be discussed in Chapter 3. Table 2.6 provides a summary of previous studies in the literature on investigating the influences on users' performance in usability testing and testing outcomes, and it shows differences between the studies.

2.3.3 Distraction

Some previous studies did not provide an exact definition of distraction, while others attempted to describe it precisely. For example, Trafton et al. (2003) described distraction as the "anatomy of an interruption". A few other studies attempted to develop a framework (e.g., Speier et al., 2003). However, the research on interruption and multitasking is currently inconsistent because of the lack of consistency in the definitions and concepts used in the literature.

Previous studies provided several different meanings and/or descriptions of terms. For example, based on Trafton et al.'s (2003) model, an interruption was defined as an alert for a secondary task (Chisholm et al., 2001; Czerwinski et al., 2004), the underlying secondary task (Li et al., 2012) or the entire pattern represented in Figure 2.2. Inconsistencies also exist in definitions of multitasking, such as concurrent multitasking (or dual task performance), interleaved multitasking (or task-switching) and sequential multitasking (Loukopoulos et al., 2009). These definitions, however, were formalised to represent different positions on a continuum depending on the task-switching rate (Salvucci et al., 2009).

Table 2.6. Comparison of Studies Investigating Influences on Users' Performance in Usability Testing/Testing Outcomes

		Empirical Comparative study	**Same data collection method**	**Type of environment**	**Adopted usability testing for the non-lab environment**		**Participants**		**Contextual factor gathered**	**Relationship between the outcomes and the contextual factor investigated**	**Conclusion about the source of difference**	**Validity**
					Asynchronous or synchronous	Formative or summative	**Type**	**Sample size**				
Andrzejczak and Liu (2010)		√	×	Lab and Field (both adopted in the university lab rooms)	Synchronous	Both	Adults	60 (30:30)	×	×	×	External (questionable)
Khanum and Trivedi (2010)		√	√	Lab and Field (both adopted in the school rooms)	Asynchronous	Formative	Children	18 (9:9)	×	×	×	External
Greifeneder (2011)		√	√	Lab and NE	Asynchronous	Summative	Adults	31 (13:18)	√	√	×	External
The proposed methodological approach in this thesis	**Study 1**	√	√	Asynchronous	Lab and NE	Predominate Summative + Formative	Adults	30 (10:20)	√	×	×	External
	Study 2	√	√	Asynchronous	Lab and NE	Predominate Summative + Formative	Adults	96 (48:48)	√	√	×	External
	Study 3	√	√	Asynchronous	Lab and NE	Summative	Adults	48	√	√	√	Internal

Another issue is that externally triggered task-switching has sometimes been called multitasking in the experimental literature (e.g., Katidioti and Taatgen, 2014), while it has commonly been called interruption in the health care literature. That is, a single definition was deemed possible and desirable, as assumed by McFarlane (1997), for example.

Regarding better research practice, suggestions have been made for future observational studies regarding the definition of distraction, which can be summarised as follows: First, definitions should be formalised according to the context and the research hypotheses or questions. Second, they should be formalised precisely to reduce error or/and bias. Some researchers have supported the concept of a universal definition, such as Brixey et al. (2007), Grundgeiger and Sanderson (2009) and Sasangohar et al. (2012). However, if it is possible, such a definition needs to be formalised or redefined each time it is used in a new context, which contradicts the purpose of a universal definition. Another important issue to consider, especially in high-traffic environments, is that an operational definition must clearly differentiate what is to be considered an interruption or a multitasking so that observed behaviour can be recorded in a repeatable way (Hintze et al., 2002). This practice will effectively enhance the comparison of results. Additionally, defining and operationalising definitions can be tested iteratively to reach a form that has minimal bias and error (Grundgeiger and Sanderson, 2009).

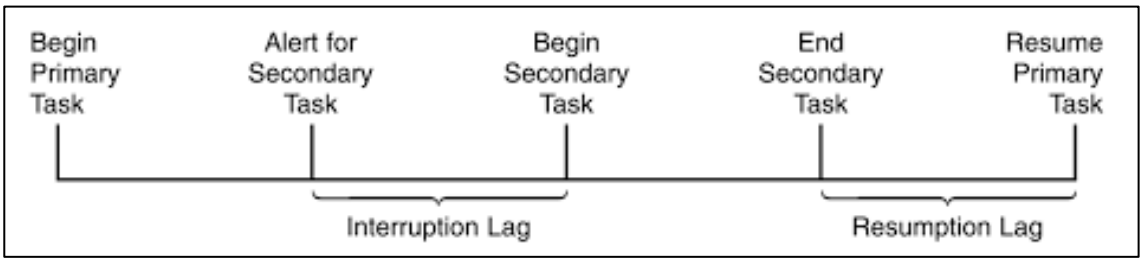


Figure 2.2: Anatomy of an interruption (Source: Trafton et al., 2003)

For the empirical work conducted in the present research, we adopted Cohen's (1980) definition of distraction and his distinction between interruption and multitasking. We believe that his definition is applicable and suitable in usability testing contexts based on the exploratory study described and discussed in Chapter 4. Cohen (1980) defined interruptions as uncontrollable, unpredictable stressors that produce information overload, thus requiring additional effort. Interruptions typically "require immediate attention" and "insist on action" (Covey, 1989, pp. 150-152). In other words, the timing of the occurrence of interruptions made by persons, in events, or by objects is beyond control. Furthermore, an interruption

breaks the attention to the primary task and forces it toward the interruption—if only temporarily. Both interruption and multitasking can occur during the performance of a primary task. However, they are perceived differently through the individual's sensory channels. In multitasking, the individual uses different sensory channels in the primary task, which may be ignored or processed concurrently with the primary task (Cohen, 1980; Groff et al., 1983). Interruptions, however, use the same sensory channels as the primary task. If the individual does not interrupt the task, he/she definitely cannot choose to ignore the interruption cues, which causes both capacity and structural interference (Kahneman, 1973). This distinction between interruptions and multitasking necessarily leads to the discussion of the sources and cost of distraction, which will be discussed later.

In addition, this empirical research adopts Trafton et al.'s (2003) model, which identifies four critical events in describing an interruption (see Figure 2.2). Prior to responding to the interrupting task, an alert could draw attention to the forthcoming event. Such an alert may provide essential information, such as urgency, which would help in deciding when and how to respond to the interrupting task (Altmann and Trafton, 2004). For example, a phone ringing alert would draw attention to the interrupting task and the phone call, and hence the decision of whether to write notes on the current task or terminate it. The other three events are the interrupting task, the end of interrupting task and the resubmission of the primary task.

"Interruption lag" refers to the time taken between the alert and the actual start of the interrupting task. An interruption lag is helpful in recording information related to the suspended primary task, which is essentially an interrupted position. The findings of several empirical studies have suggested that an alert is helpful for the resumption of the primary task (Adamczyk and Bailey, 2004; Altmann and Trafton, 2007). Furthermore, McFarlane and Latorella (2002) and Trafton et al. (2003) indicated that an insufficient interruption lag impairs the performance of the primary task. The term "resumption lag" refers to the length of time between the end of the interrupting task and the resumption of the primary task. This time is utilised to recall the interrupted task through memory or physical clues (e.g., the position when the interrupting task has taken place).

Interruptions originate from different sources. Czerwinski et al. (2004) proposed that more than half of interruptions are initiated by environmental cues, such as a new task (19%) and a telephone call (14%); the remaining are self-initiated interruptions (40%). This

classification is adopted in this thesis. However, other frameworks have been proposed to categorise self-initiated interruptions. Examples are Beeftink et al. (2008) and Jin and Dabbish (2009) (see Table 2.6). Typical examples of external interruptions are receiving phone calls, receiving emails and in-person conversations, as shown in Czerwinski et al. (2004). With the exception of Lee and Duffy's (2015) categorisation framework of external interruptions and cognitive and motor interruptions, no well-established taxonomy of external interruptions has been proposed. However, in observational studies, the categorisation of external interruptions has traditionally been adopted with respect to the specific work area, nature, or interest in the underlying work (Grundgeiger and Sanderson, 2009). In general, various types and forms of distractions are unlikely to have either equivalent influences on decision making (Speier et al., 2003) or equal negative consequences (Atchley and Chan, 2011; Sasangohar et al., 2012).

The majority of findings in the literature suggested that distractions lead to increased errors in procedural tasks (e.g., Gupta et al., 2013; Li et al., 2008), problem solving tasks (e.g., Adamczyk and Bailey, 2004; Speier et al., 2003) and decision making (Croskerry, 2013; Speier et al., 2003). Other findings suggested that distractions have disruptive effects, such as increased error rates (Li et al., 2008; Westbrook et al., 2010), difficulty in resuming original tasks (Mark et al., 2012; Monk et al., 2008; Westbrook et al., 2010) and increased feelings of stress and frustration (Mark et al., 2008).

Table 2.7. Proposed Categorisation of Self-Initiated and External Interruptions

	Interruption classification	**Proposed framework**
Beeftink et al. (2008)	Self-initiated	Self-initiated breaks Daydreaming Spontaneous or instructive thoughts Thinking about something else due to trigger
Jin and Dabbish (2009)		Adjustment Break Inquiry Recollection Routine Trigger Wait
Lee and Duffy (2015)	Initiated by others (person(s) or environment)	Cognitive Motor

Based on Cohen's (1980) definition, interruptions are likely to lead to the loss of memory or confusion regarding the information cues residing in memory, thus negatively influencing performance (Laird et al., 1983). The reason is that interruptions lead to both capacity and structural interference (Kahneman, 1973). Capacity interference occurs when the number of incoming cues is greater than the decision maker can process. Structural interference occurs when the decision maker must attend to two inputs that require the same psychological mechanisms (e.g., computer-digital tasks and in-person conversations). That is, the decision maker must respond to interruptions while performing some other activity. As result, these circumstances can place greater demands on cognitive processing resources than those available (Norman and Bobrow, 1975), likely causing loss or confusion in memory content or cues and ultimately negatively influencing performance (Laird et al., 1983). The resumption lag indicates that an individual would need more time and effort to resume the primary task after an interruption. However, if a person intentionally spends more time on recalling or planning the primary task after an interruption, performance is increased in terms of the resumption and execution of the primary task (Brumby et al., 2013).

Studies on distraction can be classified into three categories: observational studies, controlled experimental studies, and computer simulation studies (Shadish et al., 2002). Observational studies seek to detect distraction events and investigate how work/task performance will be influenced in the actual working environment. This realistic design achieves a high level of internal validity and results in generalisability. Experimental studies and computer studies, however, mainly seek to investigate the effect or the cost of distractions on work task performance or practice. That is, they are designed to control known and unknown sources of bias and thus achieve a high level of internal validity. However, they might lack adequate external validity (Shadish et al., 2002), and the generalisation of the results might be highly dependent on the extent of similarity between the study design and the actual workflow setting.

Observational studies can help to gain insights into behaviour, interactions, individual motivations and psychological processes. These factors might be crucial in studying complex socio-technical settings. For example, Nugus and Braithwaite (2010) used an ethnographic approach to address issues that might decrease the quality of organisational efficiency, including multitasking and interruptions. In their study, Colligan and Bass (2012) adopted both direct observations and semi-structured interviews.

With respect to experimental studies, some aimed to reproduce interruptions or multitasking in the context of interest, such as an office environment (Mark et al., 2008), an operating room (Liu et al., 2009) and a motor vehicle (Watson and Strayer, 2010). However, in complex and unpredictable settings, such as hospital emergency departments, such replications would become highly difficult. In complex scenarios, computer simulation studies have sought to model interruptions or multitasking in a controlled way (e.g., Lebiere et al., 2001; Sierhuis et al., 2007). The limitation of this approach is that it is highly dependent on the accuracy of assumptions. In addition, in controlled experiments, it might be difficult to capture uncontrolled environmental complexities.

The above discussion showed that it is difficult to obtain a complete picture of the environmental influences in a single study. It is conceivable that in order to gain deep insights, both approaches should be utilised. Therefore, the methodological approach used in this thesis is designed to use both approaches—observational and experimental.

2.4 Summary

This chapter has presented background information about usability and reviewed its common definitions. The definitions in Nielsen (1993) and ISO9241-11 (1998) have been considered the basis for achieving the usability of a website (Downing and Liu, 2011). Models that are characterised by similar or different usability attributes were also reviewed (Preece et al., 1994; Shneiderman and Plaisant, 2005; Quesenbery, 2001). The literature review provided in this chapter included previous studies on the perspectives, types and categorisations of UEMs, as well as usability testing methods. The formative and summative approaches to usability testing were discussed. The variants in usability testing were discussed and categorised based on technical aspects and testing locations. In addition, factors that have been found to influence usability testing were presented and discussed. Factors related to the testing environment were reviewed, as well as models with respect to context (Nilsson and Siponen, 2005; Virzi et al., 1996; Elliot et al., 2004), including the 4FFCF model. In particular, the 4FFCF model considers the factor of environmental influence, which is related to the questions addressed in the present research. This chapter also critically reviewed prior studies on RAUT, describing how they were designed and discussing overlooked validity and environmental factors. In addition, the few studies that attempted to

investigate the influence of the testing environment on usability testing outcomes were reviewed with regard to the knowledge gap that this research aims to fill.

In reviewing the literature, it was found that distraction was found to be the most influential environmental factor on users' performance and hence usability testing outcomes in the present research. That is, the anatomy, definitions and elements of distraction were reviewed. The models adopted by this research to formalise distraction and characterise it (Cohen, 1980; Trafton et al. 2003) were presented and discussed. The sources, types, influence and cost of distraction were also presented and discussed. The literature on distraction was reviewed and discussed. In the next chapter, we will provide a detailed description of the methodological approach adopted in this thesis.

Chapter 3: Methodology

3.1 Overview

Research in the field of HCI requires a methodology that will provide in-depth understanding and knowledge (Lazar et al., 2010). Creswell and Plano Clark (2017) defined methodology as the overall process or model applied by the researcher to conduct a study and fulfil pre-defined research objectives. The research methodology can therefore be regarded as the overall blueprint of a study as well as the various components of that blueprint. To choose the most appropriate research methodology and to "safeguard against making elementary errors" (Denscombe, 2003, p. 1), researchers must examine the available research methods, techniques and designs.

Following the introduction to the research and the literature review in Chapters 1 and 2, respectively, this chapter aims to justify the choice of research methodology through a general discussion of the underlying research paradigm and a description of the main research method and design used in the study. The chapter then discusses the factors considered in the experimental design phase, the methodological techniques used to collect the empirical data, and the strategies used in the data analysis. Lastly, it describes the present research design based on the formulated theoretical framework and the rationale for the methodology applied in this research.

3.2 Research Paradigm

The term research has been defined as "investigation or experimentation aimed at the discovery and interpretation of facts and revision of accepted theories or laws in light of new facts" (MacKenzie, 2013). The overall approach that guides the research and the techniques, methods and strategies used to acquire the knowledge required (Ernest, 1994) is called the "research methodology". All research is based on assumptions of how the world is perceived and how we can best come to understand it. These assumptions provide the justification for the research's theoretical stance (Creswell, 2013) and hence its methodology (Flick, 1998). In the research community, this "basic set of beliefs that guides actions" (Guba and Lincoln,

1994, p. 17) is referred to as a research philosophy or paradigm* (Lincoln et al., 2011; Mertens, 2010). It is important for the researcher to understand the philosophy adopted for the study (Tashakkori and Teddlie, 1998) because it involves important assumptions based on which the researcher views the nature of science (Saunders et al., 2009).

Researchers develop paradigms based on their discipline orientations, research communities, past research experiences, and research objectives and goals. Based on the beliefs and aforementioned factors, researchers adopt a strong quantitative, qualitative or mixed-methods approach in conducting their research. Four widely discussed paradigms are post-positivism, constructivism, transformative, and pragmatism. The elements of these paradigms differ, which is reflected in philosophical assumptions in terms of ontology ("What is the nature of reality?"), epistemology ("What is the relationship between the researcher and that being researched?"), axiology ("What is the role of values?"), methodology ("What is the process of research?"), and rhetoric ("What is the language of research?") (Creswell, 2013, p. 13). Although there has been an ongoing debate about the paradigms that researchers bring to their inquiry, answering the aforementioned questions in considering the research objectives and the elements associated with each paradigm helps to identify the desired paradigm(s) (Table 3.1).

Table 3.1: The Four Paradigms and Their Elements (Source: Adapted from Creswell, 2013)

Paradigms	**Postpositivist paradigms**	**Constructive paradigms**	**Transformative paradigms**	**Pragmatistic paradigms**
Elements	Determination	Understanding	Political and activist	Consequences of actions
	Reductionism	Multiple participants meanings	Empowerment, human rights, social justice oriented	Problem-centred
	Empirical observation and measurement	Social and heuristic construction	Collaborative	Pluralistic
	Theory verification	Theory generation	Change, emancipatory oriented	Real-world practice oriented

Crotty (1998) stated that these paradigms provide a general philosophical orientation in research, which can be combined or used individually. Even though many scholars have

* They are also referred to as "paradigms", epistemologies and ontologies (Crotty, 1998) or as broadly conceived research methodologies (Neuman, 2009).

emphasised the importance of specifying a paradigmatic standpoint that is either positivist or interpretivist, there are circumstances in which both paradigms can be combined (Gable, 1994; Lee, 1991). Indeed, some authors have called for a combination of positivism and interpretivism in the study of social phenomena to improve the quality of research (e.g., Hirschheim, 1985; Kaplan and Duchon, 1988). This assumption, otherwise termed "pragmatism", stems from ongoing debates regarding quantitative and qualitative paradigms (Tashakkori and Teddloe, 1998). The pragmatic paradigm is problem-centred and specifically considers the consequences of actions and their role in real-world practice (Creswell, 2003). Furthermore, the pragmatic approach emphasises shared meaning and joint action, reminding us that our values are always a part of our research (Morgan, 2007).

This research is based on a pragmatist view, which is the philosophical perspective suited to the research aims and questions set out in Chapter 1. As Morgan (2007) reminded us, our values are a part of our research. Although our perspective is primarily pragmatic, the emphasis of this thesis is on the postpositivist perspective rather than constructivist because the participants' performances were measured and quantified in an objective manner, and the participants' self-reports "correctly" described the world as it exists. Nevertheless, this empirical investigation incorporates some constructivist aspects regarding where the participants report their perceptions. Typically, in empirical research, a postpositivist orientation often shapes the empirical investigation and dominates the design. Consequently, it also shapes the qualitative component (Creswell and Plano Clark, 2017). Qualitative, subjective data support a better understanding of the issues under study. The pragmatic view, which implies combining qualitative and quantitative data through what is known as "mixed modes research" or "triangulation", serves to generate a broader picture of the phenomena at hand to enable the validation of research findings and to remedy the limitations inherent in a paradigm data collection technique (Creswell, 2013). Consequently, the chosen research paradigm informs the theoretical stance, which then informs the methodology used, and therefore the methods, techniques or procedures used to gather, analyse and interpret the data (Bryman 2003; Creswell and Plano Clark 2017).

3.3 Research Approach

Research approaches are the "plans and the procedural steps for research that range from broad assumptions to detailed methods of data collection, analysis, and interpretation"

(Creswell and Plano Clark, 2017, p. 3). Formalising such plans requires several decisions regarding which approach should be used to conduct the study. Such decisions help to formalise the research paradigm and research design, as well as methods of data collection, analysis, and interpretation. Because the research approach informs the research paradigm, its selection is necessarily based on the nature of the research problem, the researcher's experience and the audience of the study.

Bell et al (2018) identified two major approaches to research: the quantitative approach and the qualitative approach. According to Creswell and Creswell (2018), a study tends to be more quantitative than qualitative or vice versa. That is, the quantitative and the qualitative approaches should not be considered dichotomies because they characterise two sides of a continuum (Creswell, 2013; Newman et al., 1998). Hence, mixed-methods research falls in the middle of this continuum because it integrates elements of both approaches. The difference between qualitative and quantitative research has often been acknowledged as the qualitative framed by using words rather than numbers or closed-ended questions and responses (Creswell, 2013).

From an analytical perspective, a research approach can be outlined as deductive or inductive, which are generally associated with quantitative and qualitative approaches, respectively. In quantitative research, the researcher begins with a general review or/and observation and then involves more specific observations of the research results. That is, based on the findings of the literature review or a pre-existing theory, the researcher deduces possible explanations (i.e., hypotheses) to be tested. In contrast, in qualitative research, the researcher uses an inductive approach to plan the research. The researcher focuses on specific observations that are used to develop a final theory or conclusion (Bryman and Bell, 2011).

The quantitative approach is situated in positivist philosophy, in which a broad range of social phenomena, such as feelings and subjective viewpoints, can be investigated. Its effectiveness is highly increased when data are effectively measured and collected using the quantitative technique when a large number of data scores are available and when statistical analyses can be used (May, 2011; Goddard and Melville, 2004). In contrast, the qualitative approach is informed by the constructive paradigm (Bryman and Bell, 2011). This approach aims to investigate how the respondents interpret their own reality (Bryman and Bell, 2011).

Qualitative research is typically used to investigate the meaning of social phenomena rather than seek a causative relationship between established variables (Feilzer, 2010).

Relying on a single research approach, either quantitative or qualitative, in the postpositivist paradigm is fairly unlikely (Hirschheim, 1992). In other words, the philosophy of post-positivism suggests using mixed research techniques, including quantitative and qualitative methods (Godfrey and Hill, 1995). Mixed-methods research has been defined as "research in which the investigator collects and analyses data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or program of inquiry" (Tashakkori and Creswell, 2017, p. 4). A quantitative research study examines the relationship between variables to deductively test a theory from the literature (Flick, 1998), and the results of using this approach provide fewer details about users' attitudes and behaviours (Scandura and Williams, 2000). Thus, using mixed research methods helps obtain details and provides insights into the phenomena at hand (Punch, 2005). In the current research, a mixed-methods approach was adopted in which the researcher primarily used quantitative techniques, but applied qualitative techniques to generate a broader picture of the investigated factors and to enable the validation of the research findings.

3.4 Research Strategy

The research strategy is "a road map, an overall plan for undertaking a systematic exploration of the phenomenon of interest" (Marshall and Rossman, 1999, p. 61). The research strategy can include several research methodologies, methods and techniques. Research methods can be defined as the strategies for conducting an investigation of the phenomenon of interest, while techniques or instruments can be described as the specific means chosen to collect data (Marshall and Rossman, 1999). In the field of HCI, there are three common research strategies or methods: the observational method, the correlational method and the experimental method.

The observational method incorporates a collection of common techniques used in HCI research, including interviews, focus groups, field investigations, walkthroughs, case studies, contextual inquiries, think-aloud protocol, storytelling, and cultural probes (MacKenzie, 2013). This approach tends to be qualitative rather than quantitative, and it is used to gather information about the characteristics of the research subject without

manipulating any settings or variables (Lazar et al., 2010). Using this approach, the researcher examines and records the quality of interactions and seeks to explore and explain the reasons underlying human behaviour rather than quantifying it (MacKenzie, 2013). As a result, observational methods achieve relevance but lack precision (Sheskin, 2011, p. 67).

In the experimental method, the researcher applies controlled experiments that are typically conducted in laboratory settings either to acquire new knowledge or verify, refute, or correct existing knowledge.

The controlled setting inherent in the experimental method results in precision because extraneous factors in the real world are reduced or eliminated. A controlled experiment requires at least two variables: *a manipulated*[*] variable and *a response*[†] variable. At least two configurations are required for the manipulated variable. In HCI, a system or design often undergoes a practical "usability evaluation" or "user testing". However, these evaluations or tests do not follow the experimental method, as there is no manipulated variable. However, in a "user study", a controlled experiment is conducted in which different configurations of a variable are tested and compared. Hence, a practical usability evaluation might qualify as research; that is, information is collected about a particular subject, but it does not qualify as experimental research.

Correlational methods involve looking for relationships between variables. They are characterised by quantification because the magnitude of the variables must be ascertained. The data may be collected through a variety of methods, such as observation, interviews, online surveys, questionnaires or measurement. They usually accompany experimental methods if questionnaires are included in the experimental procedures.

Correlational methods provide a balance between relevance and precision, as data are collected using informal techniques, which brings relevance and connection to real-life experiences. However, precision is sacrificed because such methods are not controlled. In HCI research, the experimental method often includes observational and correlational methods, which is the case in the experimental method adopted in the third study in this research.

[*] Also called *independent* variable, *experimental condition* or *factor*.
[†] Also called *dependent* variable *or outcome*.

3.5 The Present Empirical Research Design

Research design can be thought of as the *structure* of research. It is "the fundamental plan of a piece of research, which contains major ideas of the research, such as the framework of the research, and presents which tools and procedures the researcher will use to collect and analyse the research data" (Punch, 2005). Research designs are "types of inquiry within qualitative, quantitative, and mixed-methods approaches that provide specific direction for procedures in a research study" (Creswell, 2013). Others have called research designs s*trategies of inquires* (Denzin and Lincoln, 2011). Research design should include all the research procedures from the problem definition to the presentation of the results (Punch, 2005). Figure 3.1 illustrates the design of the present research, including the essential steps and phases from the research problem foundation and its formalisation to the conclusions.

3.5.1 Research Theoretical Framework

In addition to the benefits of controlled contextual factors, which can be ensured before or during usability testing (e.g., type of apparatus used), other factors that are difficult or impossible to control can be explained by collecting relevant data to aid in analysing and interpreting the testing results. In this context, this means that applying usability testing with remote users in a natural environment includes the risk of exposure to distractions, such as phone calls, which can influence testing outcomes. Brewer and Crano (2000) stated, "the researchers were not only helpless to prevent such events but would not have been aware of them if they did take place" (p. 14), referring to the realisation that because disruptions can occur in a natural environment, they should be included in the data collection process.

That is, the validity of comparisons conducted using data collected from natural environments and controlled environments are more significantly influenced if distractions occur but remain unknown to the researchers. The following describes the theoretical bases that we considered in designing the approach to this research and the studies included in it.

- **Social facilitation theory**

The social facilitation theory assumes that people act differently in the presence of others than they do when they are alone. Allport (1920) coined the term social facilitation to refer to a clearly defined effect in which the mere presence of others leads to individuals' improved performance of an easy, well-rehearsed or familiar task and leads to deteriorating

their performance in complex or poorly rehearsed tasks (Fraser et al., 2001). Other researchers (e.g., Manstead and Semin, 1980; Baron, 1986) rejected this notion, believing instead that social facilitation may occur because some individuals are more vulnerable to social influences or distractions and the subsequent narrowing of attention.

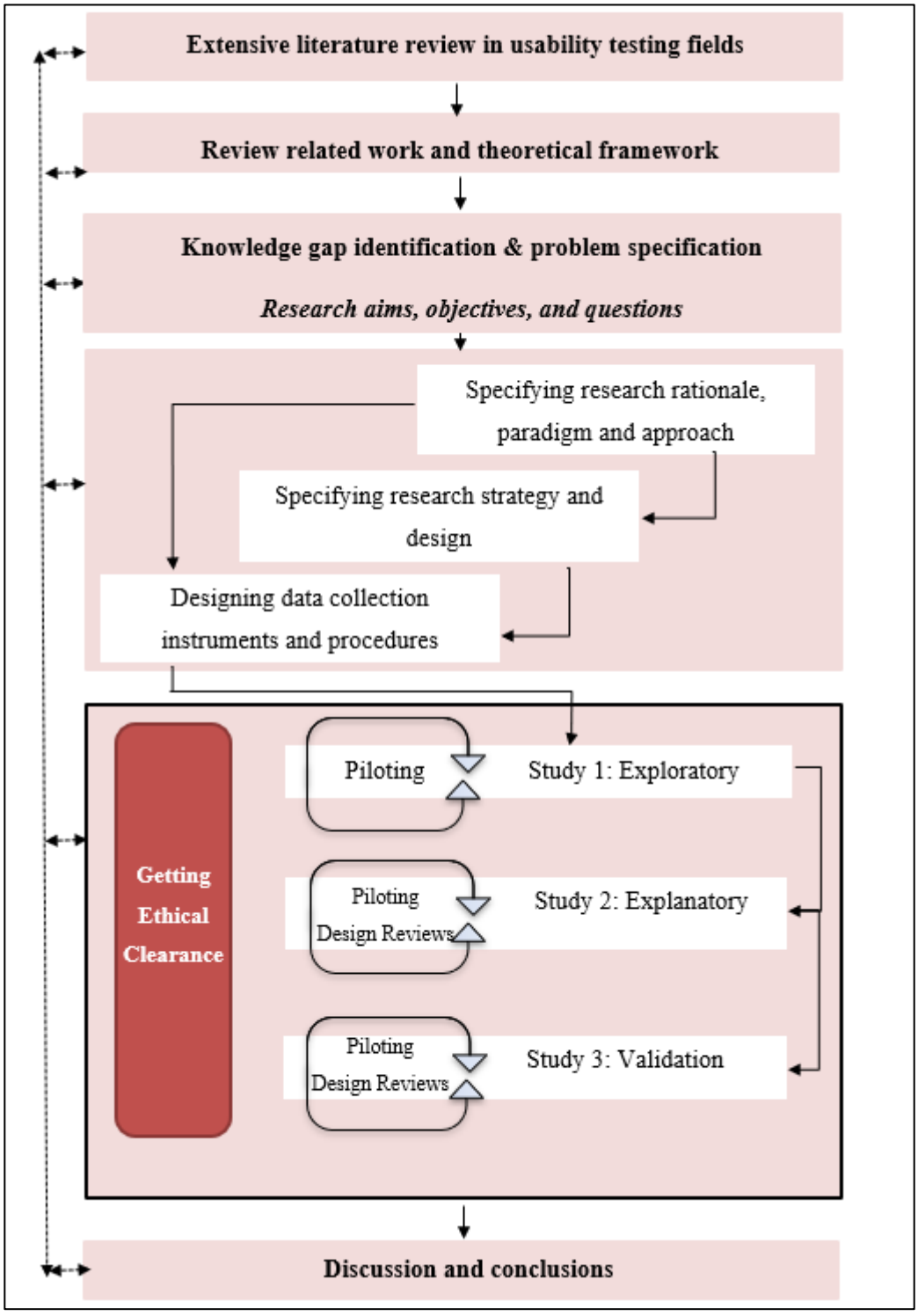


Figure 3.1: Overview of research design

These researchers argued that personality factors can make individuals more aware of others' evaluations, which hinders their performance of poorly learned or difficult tasks but does not affect or improve the performance of well-learned or easy tasks (Baron, 1986; Manstead and Semin, 1980). Social presence might also stimulate concerns about self-presentation (Bond, 1982), which might increase the cognitive effort[*] required to perform a task and therefore improve the performance of easy tasks to avoid failure and social embarrassment, which is not the case in difficult tasks (Fraser et al., 2001). Social facilitation can occur in any environment and usually when the participant is in the presence of others.

- **Distraction–conflict theory**

The distraction–conflict theory assumes that distractions do not result in amplified arousal but in cognitive overload[†], during which individuals' performances degrade in complex tasks and improve in simple tasks. Distractions help individuals make decisions by causing them to concentrate on a small number of information cues related to a simple task, leading to quicker completion times and little or no loss in decision-making performance (Baron, 1986), which is a fundamental premise of distraction–conflict theory. Performance degrades in complex tasks because the individual needs to pay attention to the stimuli related to the complex task but instead has difficulty handling all the information presented by the distractor and the complexity of the task (Bernd, 2002). The degradation effect of distractions on decision-making is caused by cognitive resources being rationed across more than one task, which eventually changes the way tasks are processed (March, 1994) and the way information is used (Baron, 1986). This, in turn, can reduce task accuracy (Cellier and Eyrolle, 1992) and cause the individual to require more time to determine solutions to problems (Schiffman and Greist-Bousquet, 1992).

- **Information overload**

Speier et al. (1999) stated, "information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur" (Speier et al., 1999, p. 338). Information overload

[*] Cognitive effort is defined as "the engaged proportion of limited-capacity central processing" (Tyler et al., 1979).

[†] Cognitive load refers to the total amount of mental effort used in the working memory. Cognitive load theory was developed to study problem solving by John Sweller in the late 1980s (Sweller, 1988).

has been found to hinder the quality of decisions (Chewning and Harrell, 1990; Snowball, 1980) by increasing the time needed to make a decision as well as misunderstanding and confusion concerning the decision (Cohen, 1980; Malhotra et al., 1982).

The most commonly cited cause of information overload is the number of information cues (Evaristo et al., 1995). In addition, the task demand, such as the task complexity level, can directly affect the mental workload required to complete the task and lead to information overload (Hart, 1986), thus affecting the decision that is made.

3.5.2 Research Methodology Rationale

The rationale for the present research was derived from reviewing and comprehending the aforementioned theories (section 3.5.1). Based on the rationale, the methodological approach and measurements adopted in this research are justified.

The research rationale process was formalised in several steps. In the first step, the relevant literature was reviewed and critically analysed. Secondly, the limitations of previous research-related knowledge gaps were identified. Thirdly, the relevant theories in the literature were reviewed, synthesised and analysed to find possible explanations for the limitations in the previous research. Fourthly, the methodological aspects related to the research area were reviewed to realise the possibility of filling the knowledge gaps in the literature. Fifthly, the research problem and the research questions were formalised. Sixthly, the methodological approach and perspective(s) were decided.

The procedures described in the first step to the fifth step, with the exception of the third step, were presented and discussed in Chapters 1 and 2. They can be summarised in the following issues that should be considered in formulating the research rationale:

- Mitigating the validity issues acknowledged in the relevant literature in terms of the limitations in the statistical tests applied, the conclusions passed to the practitioners and researchers, and the instrumentations and measures used for comparison(s).
- Ensuring the reliability of comparisons by avoiding the evaluator effect.
- Ensuring proper assessment of the usability testing practical utility by focusing on the design impact and users' feedback on usability.

- Ensuring the validity of the comparisons, instrumentation and generalisability of results.
- Applying a new approach to investigate the capabilities and shortcomings of usability testing with remote users.
- Considering contextual factors and their potential effects on usability testing.
- Ensuring the awareness of the possibility of the existence of contextual factors, their types and frequency and considering their possible relationships to usability testing outcomes.
- Exploring and investigating the source of the inconsistencies in the results reported as RAUT's outcomes in the literature compared to traditional usability testing.
- Determining whether inconsistencies in the results reported by RAUT's outcomes in the literature compared to the traditional usability testing were related to the usability testing methods used or to the testing environment utilised.
- Investigating the implications of the existence of influential contextual factors during usability testing performance for its outcomes.

The sixth step was discussed at the beginning of this chapter (Chapter 3, sections 3.2-3.4). With regard to the third step, we reviewed, synthesised and analysed the relevant theories in the distraction and work-overload literature, and we mapped their elements to elements in the 4FFCF model. Mapping helped the researcher to better realise the possible contextual factors that might take place in users' natural environments and their potential influences and implications. Table 3.2 illustrates the mapping process. The aforementioned issues led to the perception that the empirical approach should be applied to the present research.

In the fourth step, the review of the primary concepts of empirical research indicated that they were regarded as the capability of being verified or disproved by observation or experiment. Empirical research can be observational, correlational and/or experimental (MacKenzie, 2013).

In the HCI field, experiments are focussed on the interactions between humans and computing technology. Studying such interactions involves addressing their qualities, which is typically outside the scope of solo experimental procedures. Looking for and finding a circumstantial relationship is often the first step in further research (MacKenzie, 2013). As

a result, a proper user study—that is, an experiment with human participants—involves a comprehensive understanding of interaction quality, which in our context is distraction. These qualities might not appear in significant numbers, but they cannot be ignored. Regarding this point, observational methods should be involved by soliciting comments, thoughts and opinions from the participants in HCI experiments (MacKenzie, 2013; Lazar et al., 2010)

There are two possibilities in observation: manual observation by the experimenter or investigator; passive observation by an "apparatus". Observational data reveal data patterns that require to be examined, measured, recorded, and analysed to determine "significance differences". In measurement, these data patterns yield empirical evidence.

Relationships between variables can also be observed, measured and quantified. However, these observed relationships are circumstantial, and they are typically associated with correlational research methods.

In contrast, causal relationships emerge from controlled experiments where participants are randomly selected from the target population and randomly assigned to the experimental conditions, which are also known as units, conditions or treatments (MacKenzie, 2013; Lazar et al., 2010). An experimental study usually starts with a research question or a testable research hypothesis (Lazar et al., 2010).

Based on the discussion of the experimental method, two important properties of experimental research are to be considered: internal validity and external validity. Internal validity is the extent to which an observed effect is due to the test conditions; external validity is the extent to which the experimental results are generalisable to other people and other situations; that is, experimental environments and procedures that are representative of real-world situations where the interface or technique will be used. Hence, the experimental method resembles an exercise in compromise if strict considerations of internal and external validity were adopted.

There is no remedy for the tension between internal and external validity in experimental methods, so at very least, the researcher must acknowledge the limitations. Consequently, in HCI, experimental research methods are often accompanied by observational and correlational methods so that multiple narrow testable questions that cover the range of outcomes that influence the broader untestable questions increase both types of validity.

Table 3.2. Theoretical Framework of The Research

Theory				**Mapping to 4FFCF model**				
				Condition(s)/Facto(s)				**Expected implication(s) on Usability testing outcomes**
Name	**Conditions**	**Source of influence**	**Theory suggested implications**	**Environment**		**Task scenario**	**User Characteristics**	
				Distraction type	Source			
Social facilitation theory	Presence of others × task complexity	Amplified arousal	• Improved performance of an easy task **(**Fraser et al., 2001). • No change in the performance of easy tasks (Baron, 1986; Manstead and Semin, 1980). • Deteriorate performance for complex **(**Fraser et al., 2001; Baron, 1986; Manstead and Semin, 1980).	External interruption	In-person conversation	Complexity/difficulty level of usability testing task	Attitude Personality State	Performance • Time on Tasks • Successful completions • Number of page views • Errors
Distraction–conflict theory	Distractions × task complexity	Cognitive overload	• Concentration on a small number of cues lead to improved quicker performance of an easy task (Baron, 1986). • Attention is required to be paid to the stimulus of a complex task while handling the information presented from the distracting task. (Bernd, 2002). • Change in complex task processing (March, 1994). • Reduced performance accuracy of complex tasks (Cellier and Eyrolle, 1992). • Change in use of information from complex tasks (Baron, 1986). • Longer time to solve complex tasks (Cohen, 1980; Malhotra et al., 1982).	External interruption	• In-person conversation • Phone calls • Intrusive text messages	Complexity/difficulty level of usability testing task	Competence	Performance • Time on Tasks • Successful completions • Number of page views • Errors
Information overload	Information cues, task demand, or task complexity	Limited cognitive processing capacity (Mental workload)	• Reduction in the quality of decisions made (Speier et al., 1999; Chewning and Hanell, 1990; Snowball, 1980). • Increasing the time needed to decide (Cohen, 1980; Malhotra et al., 1982). • Misunderstanding and confusion concerning the decision (Cohen, 1980; Malhotra et al., 1982).	External interruption	• In-person conversation • Phone calls • Intrusive text messages	Complexity/difficulty level of usability testing task	Attitude Personality State	Performance • Time on Tasks • Successful completions • Number of page views • Errors Perceived usability • Ratings • Reports
				Multitasking	Other opened applications			
				Poor apparatus performance	• Small display size • Low connection speed			

Considering the previous discussion and taking into account the issues in the process of formalising the rationale process, the experimental method was deemed the most suitable for the present research. Nonetheless, we conducted brainstorming for the issues discussed in Chapters 1 and 2 before attempting the sixth step. We then decided how to adopt all of them in the research design.

We decided that it would be impossible to address all issues simultaneously in one study. Related issues were grouped together, and it was decided to address them in a study with a preliminary formalisation of the related research questions, which we developed based on our perception of the appropriate and applicable research design.

In addition, we realised that we needed more than one study to address the aforementioned issues. We decided that we needed to select a data collection method that could be used in all the necessary studies. At this point, we examined the technical feasibility of the data collection method as well as its reliability, validity and utility.

In addition, we realised that we needed to be aware of the contextual data while simultaneously recording the usability testing outcomes in an objective way to make valid comparisons. Consequently, we realised that we needed to explore the data collection method, its outcomes and its capability of revealing insights into what happens in a usability testing session.

We decided to adopt a comparative design to assess whether we could formalise and design a valid comparison. At that point, we decided that we needed to adopt a two-stage design in which the exploratory findings from the first stage directed the rest of the research activities.

In step five and after multiple iterations, the methodological rationale depicted in Figure 3.2 was formulated. Three studies were proposed for the data collection: the experimental methods used in each study would be accompanied by an observational and/or correlational method, depending on the objectives of each study. This approach would enhance the internal and external validity of the study and therefore the research. After formalising the research rationale, we moved to step six to decide the research methodological approach and perspective(s). After reviewing the concepts presented in sections 3.2-3.3, the research methodical aspects were established (see Table 3.3).

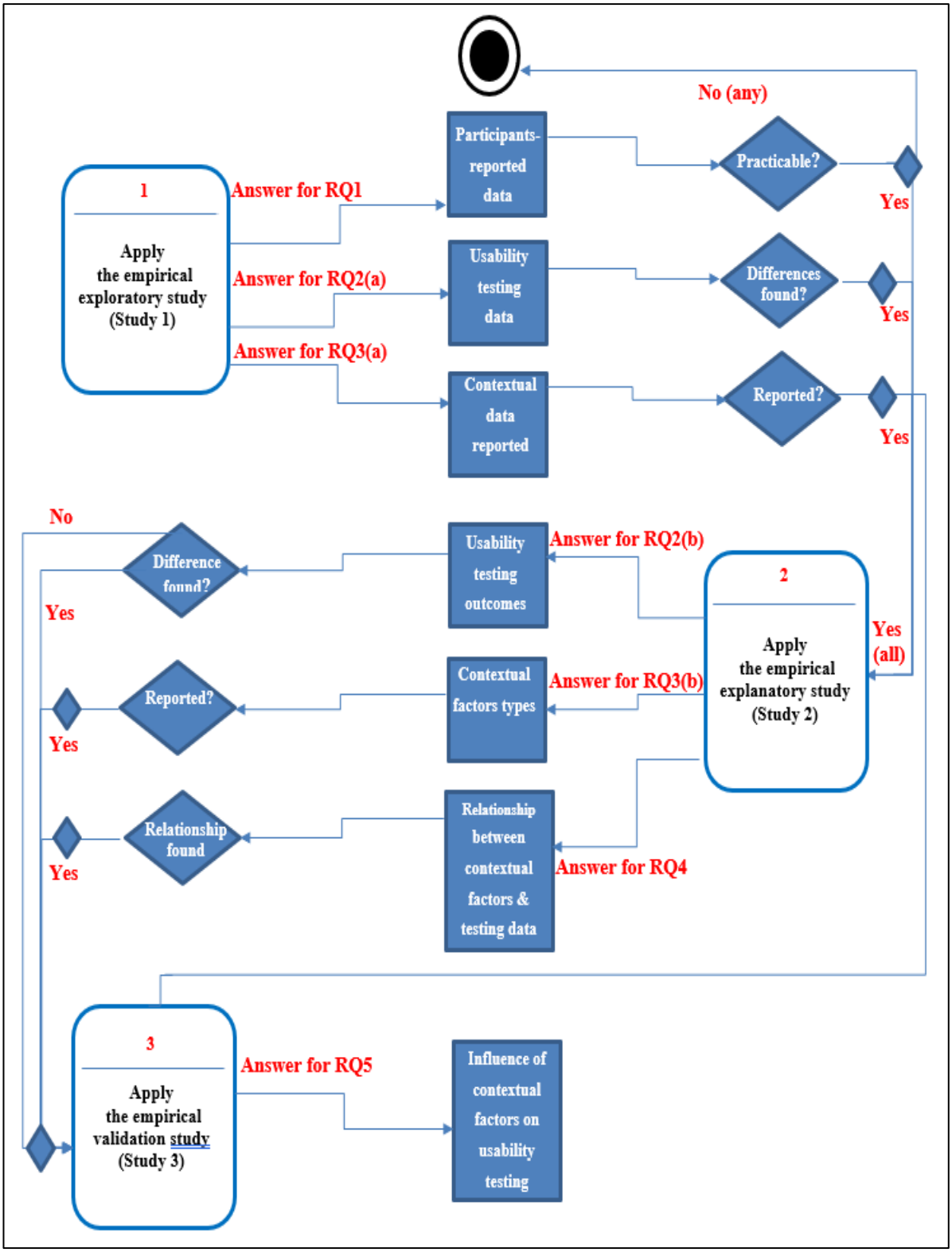


Figure 3.2. Research rationale

Table 3.3. Research Methodology

Study	Methodological approach	Purpose / objective	Research questions to be answered	Research strategy	Dominant paradigm /perspective	Data type(s) collected	Dominant research approach
Study 1	Empirical	Exploratory	RQ1, and RQ2	Comparative Observational	Postpositivist	Quantitative Qualitative	Quantitative
Study 2		Explanatory	RQ3, RQ4, and RQ5	Experimental Comparative Correlational	Postpositivist	Quantitative Qualitative	Quantitative
Study 3		Validation	RQ5	Experimental Comparative	Pragmatic (Postpositivist + Constructive)	Quantitative Qualitative	Mix-mode / Triangulation (Quantitative + Qualitative)
In general, empirical research with a pragmatic paradigm							

The methodological details of each study will be provided at the beginning of each relevant chapter. The best practices regarded in the relevant literature on empirical research are provided in Appendix A. CH4 is followed when it is relevant to each study.

3.6 Summary

This chapter has presented the justification for the empirical methodology and the approach followed in this research. The pragmatic paradigm, research strategy and type of data to be collected were described and discussed.

To answer the research questions, three studies were designed and undertaken. Study 1 was designed as an exploratory study to answer the first and second research questions. Study 2 was designed as an explanatory study to answer the second, third, fourth and fifth research questions. Study 3 was designed as a validation study to answer the fifth research question. The comparative research strategy used in all three studies. In study 1, the observational approach was based on a predominantly postpositivist perspective. Study 2 was designed using an experimental, comparative, and correlational research strategy based on a predominantly postpositivist perspective. Study 3 was designed using an experimental comparative research strategy based on a predominantly pragmatic (postpositivist + constructive) perspective. The data collected in all three studies were both quantitative and qualitative. Mixed modes triangulation was applied to both data strands in study 3 to answer the fifth research question.

Chapter 4: RAUT in Natural Environment

4.1 Overview

The previous chapter provided the methodology of this research. The review of previous research in the area of RAUT showed that the majority of the conducted studies were predominantly limited to comparing RAUT with other UEMs, typically traditional laboratory testing, with little or no awareness about what might happen during the usability testing session. In the case of RAUT, we refer mainly to contextual factors, specifically distractions. Thus, some additional work needs to be performed to explore what happens during RAUT sessions in participants' natural environment (NE).

In this chapter, we present the proposed empirical data collection method used to collect data on participants' performance and on contextual factors during the usability testing session. The aim is to use an online unmoderated usability testing (OUUT) tool to administer the usability test online so it can be accessible in any environment via the Internet. These tools guarantee the objective automatic recording and quantification of participants' performance. In addition, these tools enable the online administration of textual instructions and questions, which enable us to gain insights into what happens during the usability testing in the form of data reported by participants.

The online administrated usability study is designed and implemented, and its capability is explored to provide data on usability testing outcomes in terms of participants' performance and to obtain insights about the contextual factors that might arise. The findings are promising. Issues are raised from each testing environment and several suggestions for improvement are offered. The remainder of the chapter is organised as follows: Section 4.2 introduces the design, analysis and findings of this study; Section 4.3 discusses the findings; and Section 4.4 discusses the study limitations.

4.2 The Empirical Exploratory Study

4.2.1 Study Objectives

In this exploratory study, we explore the capability of an online usability study via usability testing and questionnaire to collect data in different environmental settings, giving consideration to the different factors related to the testing environment. This study aims to

answer the first research question (RQ) and contribute to the second and third questions:

RQ1: What can we expect from participants of remote usability testing when they are asked to report their issues and outcomes?

RQ2: Does usability testing data performance during usability testing in the (remote) natural environment differ from that of participants in a lab environment?

RQ3: What contextual factors do remote participants experience during their usability testing session?

To answer these questions, this study uses the OUUT to meet the following objectives:

- Explore the functionality of usability studies in administering the test and its tasks, instructions and questions within different experimental settings.
- Explore the data provided by participants through the online administrated usability study about the interaction with the test object(s) during the testing session in the different testing environments.
- Explore usability outcomes in different testing environment settings.

The process of designing, administering and launching the study will obtain aggregate results to ensure the data do not contain improbable values, oddities and contradictions in the success rates, ratings and comments. Performing the intended analysis on the exploratory data will provide insights into any problems with the study design. We examine the recruitment process, such as sending vouchers, dealing with participants and estimating the level of interest shown by people in participation in the study. We also practise collecting, exploring and preparing data in both environments; select analysis approaches and appropriate statistical tests for the data; and report the results.

4.2.2 Study Design

To answer the first three RQs, we need to design a tool that enables us to apply usability testing with remote users that is accessible to participants in other environmental settings. Therefore, we empirically investigate how participants perform usability testing in different environments. As indicated in Section 1.4, when this type of user study is implemented, accessed and utilised online, it is called an online user study. In our case, the user study was designed to collect data regarding website usability and participants' performance on those websites; hence, we called it an 'online usability study'.

There are multiple benefits to selecting an online usability study as a data collection method, especially in different locations, environments, situations or contexts, as they can be applied

and accessed online anywhere at any time. Online usability studies can be assigned to remote users in their NEs and with users in laboratory settings as long as there is an online connection to access the usability study. The unified data collection method and online access mechanism allow identical methodological access in different environmental settings, thus creating a control group in the experimental comparison for the differences concomitant with using different UEMs (as mentioned in Sections 1.2.1 and 1.2.2), which increases the validity in terms of instrumentation and setting.

Online usability studies can also be conducted as a combination of web testing (scenario-based testing) and surveys. Web testing imitates the scenarios or tasks given to the participants in the lab usability test and is complemented via a questionnaire(s), meaning various types of questions can be asked. The aforementioned advantages make an online usability study ideal to use so that two different groups of participants can perform usability testing tasks in different environments, i.e. in a lab and in participants' NE.

This type of online study could be designed using automated tools with no observation (unmoderated) or passive observation (indirect, moderated). The control for the evaluator (Hawthorne) effect* could be achieved for both unmoderated and moderated types if the test participants do not know they are being moderated; however, the latter method is equivalent to 'spying' on participants and is unethical. Since this thesis investigates the implications of applying usability testing with remote users in different environments and in the form of asynchronous usability testing, typically no physical synchronous direct observation is carried out by the observer. Therefore, an unmoderated online usability study was chosen.

Having selected an online usability study as the data collection method, we specified the data required to answer the relevant RQs. Since this study is exploratory, several data are collected. Based on the 4FFCF model, any typical usability testing is based on measurements for its outcomes, which are represented by performance and perceived usability measurements.

Referring to the theoretical framework specified for this research (Table 3.2), the applicable measurements for usability testing outcomes for this study were as follows:

* The reactivity in which individuals modify an aspect of their behaviour in response to their awareness of being observed (McCarney et al., 2007).

- Performance outcomes
    - Efficiency measurements
    - Effectiveness measurements
- Perceived usability
    - Subjective scores
    - Subjective reports

Besides the measurements on the usability testing outcomes, to answer the RQs, we explore data related to the test and collect data on what was happening during the test. The data shown in Figure 4.1, which lie nominally within the red box, will be called usability testing data from now on. Usability testing data represents data scores that do not belong to usability testing outcomes or to the contextual factors specified by the 4FFCF model. Such data, besides usability testing outcomes, are examined for the presence of contextual factors. In this study context, these data are represented by the time consumed by the participants to answer the questions and read instructions. Depending on the data collection capabilities provided by the OUUT used, measurements on usability testing outcomes and testing data are operationalised and specified to reflect the actual experience of the participant during the usability testing session.

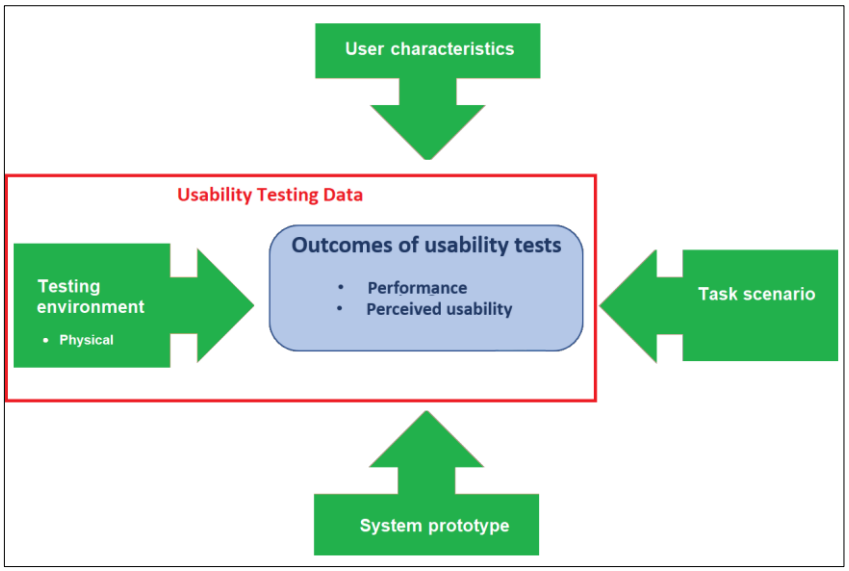


Figure 4.1. Usability testing data with respect to 4FFCF model.

4.2.2.1 OUUT Tool: Loop11

Loop 11 is an online unmoderated tool (www.loop11.com), which provides built-in test templates that enable the administration of usability testing. We we used the OUUT, Loop11

builder to design the experimental usability testing tasks to use with the chosen target website(s).

Loop11 is an affordable tool with a task-based interface which allows participants to select 'task complete', 'abandon task' or 'continue'. Users choose 'task complete' when the required information has been retrieved. However, if the information retrieved is incorrect, the task is considered a failure. Users select the 'task abandon' option when the related information cannot be found. Option 'continue' allows users to proceed to the next step in the test, mainly between questions. There was no need to install any additional software. All digital libraries were fully functional within the test window. The participants were not required to switch between windows to view the questions and the digital libraries' websites. The testing task required obvious and assessable endpoints to enable Loop11 to indicate the success or failure of the corresponding task. To accomplish this functionality, the researcher provided the URL of the target page. Accordingly, in the testing session, the tool tracked the target of the participant's navigation path and indicated whether the participant succeeded. During this study, Loop11 enabled the task to appear only at the top of the window of the website being tested. In addition, Loop11 enabled the researcher to locate the questions before or after the tasks as desired. As a result, instructions were presented before the tasks, and questions asking about the answer, usability ratings and issues and assessing them were presented after.

4.2.2.2 Experimental Usability Tasks

The aim of this formal[*] empirical study was not to evaluate the usability of particular websites but rather to investigate usability testing in terms of users' performances and their perceived usability in different environments.

As it was planned to recruit UEA students as participants for the study, digital library websites were chosen as the target websites as the participants were users of such websites. With such websites, the tasks could be formulated and designed to be similar to students' objectives when locating supporting information for essays and coursework. The tasks were a collection of predefined simple, medium and highly complex tasks. They simulated problem-solving tasks, where the evaluator had tested the website before, asked the participants to find what

[*] Formal study in this context means non-practical usability testing study; it is a research-oriented study aimed to communicate knowledge to the related research body.

they asked for and provided a hint so participants could verify whether they had found the correct answer; this was to ensure that the participants could solve the tasks (task success measurement). To avoid making users panic or feel that they were being examined, users were asked in the test whether they thought they found what was required of them instead of asking them for the correct answer.

With respect to the test object(s), there were two possible options when designing this usability study: (1) to use *one* website with four completely *different* assorted empirical *tasks* or (2) to use *similar* tasks (e.g. all searching for a resource) with *different websites*. We opted for the latter option because we believe that with digital library websites, the type of tasks to be applied are limited in functionality, as they tend to be based on or around the main search function. As a result, we argue that for better comparison and generalisation, it is better to design similar searching tasks, but with different libraries' websites. In addition, with this research, we aim to obtain more comprehensive insights about usability testing outcomes, which would also offer insights into usability issues with the test object(s). We consider that it is somewhat difficult to ask users to report usability issues after completing all the tasks on one test object. The participant might forget the issues that came to mind after completing all the tasks. Therefore, given that the nature of online usability study design does not allow for direct reporting for usability issues, as for example do UCIs, we believe it is better to involve more than one test object. Hence, the tasks were designed as searching tasks where participants search for a specific item (e.g. a file) and specific information on the website or in the retrieved resource (see Table 4.1).

Table 4.1. Experimental Tasks Purposes and Objectives

Task ID	**Task Purpose**	**Task Objective**
Training Task	Training	Search for publication date of the retrieved resource
Task 1	Actual	Search for author name and publication date of the retrieved resource
Task 2	Actual	Find the number of verses in the retrieved resource
Task 3	Actual	Find the number of figures in the retrieved resource
Task 4	Actual	Find the number of pages in the retrieved resource

The same four actual usability testing tasks were used in both environments, which all utilised Loop11, such that for every digital library one task was designed. This study used four digital libraries' websites that are freely available online: CiteSeerX, Perseus, arXiv, and JSTOR. Amazon (www.amazon.co.uk) was also selected as a control website.

With the Amazon website, participants were required to search for book(s), as this is similar

to the searching tasks in digital libraries. Amazon was used as a control website in this study because it has a permanent URL and provides relatively stable search results. It also has a relatively familiar and well-designed interface. Therefore, data yielded from Amazon tasks in the different environments could support or contradict the data yielded from the other digital library websites. The digital libraries, which were selected after investigating their specialties and interface designs, were used and tested by the researcher and were found to have several usability issues (see Table 4.2).

Table 4.2. Digital Libraries' Websites Used for The Study

Target Digital Library	**URL**	**Specialties**	**Provider**	**Corresponding Task**
JSTOR	https://www.jstor.org/	General	ITHAKA	Training Task
CiteSeerX	https://citeseerx.ist.psu.edu/	General	Pennsylvania State University	Task 1
Perseus	http://www.perseus.tufts.edu/hopper/	Arts and Humanities	Tufts University	Task 2
arXiv	http://arxiv.org/	Applied Scientific Subjects	Cornell University	Task 3
Amazon UK	http://www.amazon.co.uk/	E-commerce	Amazon Co.	Task 4

4.2.2.3 Ethical Clearance

Once the experimental materials were fully designed, all the documentation, including the required participant reassurances, screenshots of the study design materials and informed consents were submitted to the Ethical Approval Committee of the Computing Science School at UEA.

4.2.2.4 Experimental Protocol

After receiving ethical approval from the Computing Science School's Ethical Committee, we started the data collection process. Thirty participants (60% male) aged 18-33 years were recruited from UEA schools (mean = 24.23, SD = 4.2). Of the participants, 20 were recruited for the NE group and 10 were recruited for the lab group. Their education background ranged from undergraduate to PhD level.

Participants for the NE group were recruited through emails, Facebook, Twitter and advertisements on the university's school bulletin boards. All emails and messages contained an introduction to the study and a link to its web portal, which provided additional information about the test, instructions, participation consent and contact information. The direct link to the test was not provided in the initial invitation email. However, this approach

did not yield a good response rate. Therefore, we included a direct link to the study to recruit more participants for the NE group. Participants were told they could take the test at a time convenient for them within a week.

Participants for the lab group were recruited via flyers placed throughout the UEA campus. The flyers contained a brief introduction to the study and location of the testing room and testing time, which was between 9:00 AM until 5:00 PM for one week. To prevent users from choosing the usability testing location prior to the actual test, the URL to the study portal was *not* printed on the flyers; instead, it was shown on a sheet next to the computing machine where the usability study was administered. In the testing room, only the Safari browser was installed in the machine to be used for the testing. Similarly, only one participant per session was allowed to be in the testing room during the experiment to avoid distractions.

Participants in both environments were unaware of the other usability testing environment. Information pertaining to non-lab usability testing was not mentioned to the lab participants and vice versa to avoid demand characteristics response bias (Nichols and Maner, 2008). In addition, no guidelines were provided to the participants regarding multitasking or interruptions. However, participants in the lab environment were asked to avoid being distracted while carrying out the test; this was affirmed in the sheet provided beside the usability testing machine in the lab. Figure 4.2 shows an overview of the experimental protocol adopted for the exploratory study.

A web portal was designed to enable unified access to the test from anywhere including the lab and to unify typical testing procedures, such as obtaining consent from the participants (see Figure 4.3). In addition, the portal was designed and implemented to guide the participants through the testing process. This platform provided centralised, real-time support without the need for a human observer to be present. Two versions of the portal were designed: (1) for regular web access using standard versions of browsers and (2) for mobile networking access using android versions of browsers.

The portal introduced the study's usability testing website, which was designed using Loop11, and the pre-test instructions. The portal also presented contact information, frequently asked questions and the consent form (Figure 4.3). Once a participant agreed to participate in the study, they were directed to the usability testing website where further instructions were provided. If the participant declined to participate, the session terminated;

however, even those who agreed to participate could withdraw at any time during the test. If the participants agreed to participate, whether with access from the natural or lab environment, they were transferred to the unified usability testing website by Loop11, where they were briefed about the objectives of the usability testing prior to performing the tasks.

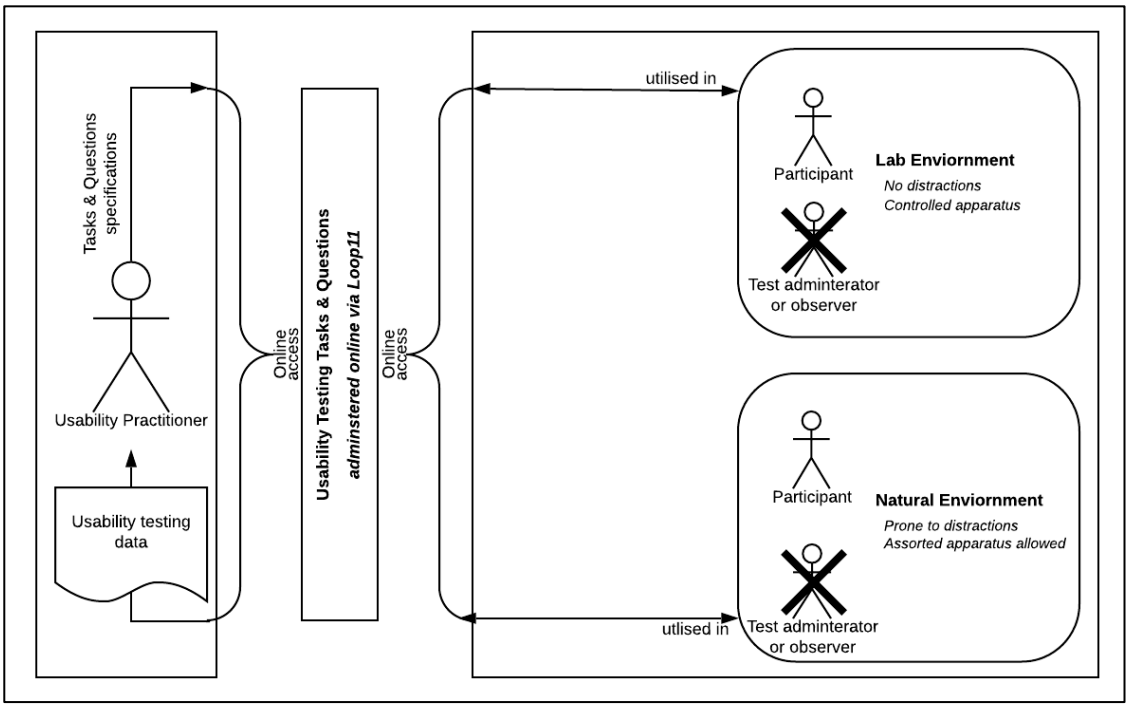


Figure 4.2: Overview of the experimental protocol for the exploratory study.

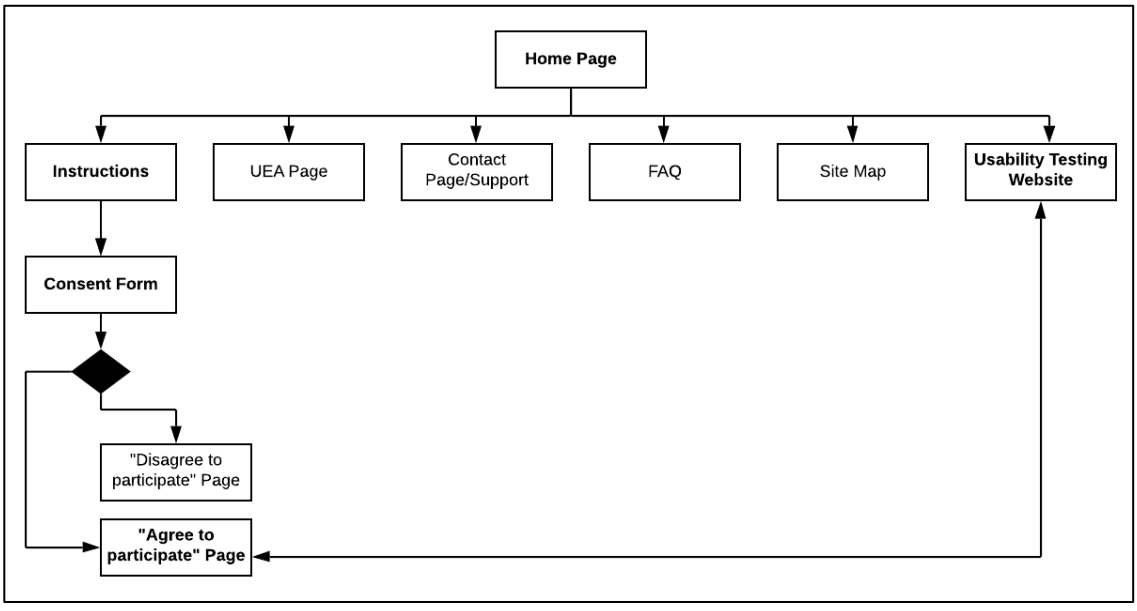


Figure 4.3: The portal website map.

Participants were asked to perform the tasks as they would normally do. They were instructed to carry out a training task before the actual test to familiarise themselves with the testing interface and the nature of the experimental tasks. We used a separate digital library

(JSTOR) for the training task, and any answers provided by the participants were not recorded.

To minimise psychological stress that may occur and to encourage participants to perform the tasks naturally, the participants were not made aware that they were being timed during the test. All participants were urged to provide honest answers to ensure accurate data and were assured that their answers would not affect their participation reward. After completing each task, participants were asked to rate the task difficulty and assess the website's usability using five-point Likert scale-based questions. They were also asked to report if they had noticed any usability issues.

Participants in both environments were asked to report whether they had other applications/programs open during task performance and whether they were multitasking. They were also asked if they had been interrupted while completing the task. If so, they were required to list these distractions and give a rating regarding the extent of the distraction caused by multitasking or interruptions. Questions pertaining to distractions, interruptions and settings were placed at the end of the usability study to avoid a demand characteristic response bias (Nichols and Maner, 2008) that may influence participants' performance of the subsequent task(s) or their answers to the subsequent question(s). For the lab group, participants were required to use a UEA machine in a specified room in the UEA library utilising UEA's standard Safari browser and network. However, the NE test participants used their own machines, browsers and network connection technology.

The study participants were given the options of voluntary participation or a £5 Amazon.co.uk voucher incentive. The participants were allowed to decide the type of participation, thereby limiting the chances of sample error due to over-motivated or profit-seeking participants. We realised that the NE test setting might be more attractive to potential participants, so to avoid reaching the limit of the available vouchers, potential participants were informed prior to consent that the vouchers would be subject to availability. Just before the end of the test, each participant was asked to provide an email address for delivering the e-voucher or to skip the email question if they opted for voluntary participation. Pilot tests are especially important for this kind of study when no moderator is physically present. The study could have poor quality results if it is not pretested (Albert et al., 2009). As a result, besides technical checks, both technical and usability checks were carried out. Loop11 (Refer to section 4.2.2.1) is technically reliable; however, technical checks were essential to check the links, data passing and branching from the portal, since it was designed and built

by the researcher. The usability checks covered both the portal and usability testing website designed by Loop11. Both checks were carried out the first time with five volunteers and the second time with two volunteers until the study design achieved a satisfactory level of technicality and usability.

4.2.3 Study Analysis

This study's analysis activities were carried out in three main stages: (1) preparing the data for use in the analysis, (2) exploring the data to obtain insights into how the data were distributed and (3) performing the analysis to answer the RQs.

4.2.3.1. Data Preparation

Data were prepared for analysis by conducting coding and quality checks, which involved ensuring that all variables scores were within possible ranges, checking and addressing missing values and outliers, and determining general themes in the data to identify strange values or typographical errors. Data were also assessed in terms of normality of distribution and heterogeneity of variances.

4.2.3.2. Data Exploration

Thirty participants aged between 18 and 33 years participated in the study (mean = 24.23, SD = 4.2); 60% of the sample were male and 63.3% were native English speakers. Of the non-native English speakers, 13.3% indicated that they could read and write and were confident speaking in English, 20.0% indicated that they could read, write and chat in English, and 3.3% indicated that they could read and write but had difficulty searching and/or writing in English. The participants were UEA students; 26.7% were PhD students, 16.7% were in their first year of study, 13.3% were in their third year of study, 10.0% were doing masters and 3.3% were in their foundation year. The majority of participants (80.0%) were familiar with Amazon.co.uk.

4.2.3.3. Analysis Approach

This study used a between-subjects statistical design, where 10 participants served as the lab group and 20 served as the NE group. We used statistical tests to compare the groups' data obtained from the continuous (interval/ratio) data recorded by Loop11 and the participants (e.g. number of distractions or usability issues). We opted for the parametric statistical

independent t-test or the Mann-Whitney test depending on the distribution of the data. For the other category (dichotomous/binary) (e.g. successful completions), we used the chi-square test.

Loop11 objectively and automatically recorded the performance measurements in terms of time measurements and successful task completion. Loop11 also tracked and recorded page view measurements based on the page URLs for each task. The number of tracked URLs formed the number of page views. The participants were asked via Loop11 about perceived usability in terms task complexity, website usability and usability issues. To answer RQ1, the participants were also asked to report the number of usability issues and to describe them. However, no qualitative analysis was applied to the descriptions of the usability issues. Loop11 also acquired data on usability testing in terms of contextual issues that participants faced and about their characteristics.

4.2.4 Study Findings

The findings are divided into three subsections. The first subsection describes participants' reported data in both environments. The second subsection reports the results of the statistical tests for differences between groups (Lab vs NE) in terms of the usability testing outcomes. The last subsection describes the type of contextual factors reported in the testing sessions.

4.2.4.1. Participants Reported Data

Participants were asked to report about the usability testing they performed. At the end of the test, they were asked if they encountered usability issues and distractions after each experimental task, and they were asked to give feedback regarding the contextual factors implied by their environment when performing the usability testing, including the systems and distractions.

- **Usability issues**

Almost all the participants (99% of NE participants and 90% of lab participants) indicated that they experienced usability issues and reported the number of occurrences and descriptions. However, the Fisher Exact test showed no significant association between the type of testing environment (Lab vs NE).

- **Distractions**

Only participants in the NE group reported the occurrence of distractions during their experimental session. All the NE participants reported and described a number of distractions.

- **Apparatus**

Participants in the NE group were asked (after completing all experimental tasks) to report on the type of computing systems they used to perform the experimental tasks and the network. The participants in the lab environment were asked to confirm that they used the UEA computer and the NW provided.

4.2.4.2. Usability Testing Outcomes

Usability testing outcomes are presented as performance outcomes and perceived usability reports. The findings for the components of each outcome are detailed as follows:

- **Performance**

The data collected on performance for this exploratory study were Time on Tasks, Page Views, and Successful Task Completion.

  - Time on Tasks

    The descriptive data presented in Table 4.3 shows that the mean values for the Time on Task in the NE environment are larger than those for the lab environment. However, the Mann-Whitney U test shows that no significant difference exists for Time on Tasks (for Tasks 1-4), and for Time on All Tasks between the testing environments (lab vs NE), U = 149, z = 3.5, p = 0.432, r = 0.7, and the effect size r is considered a medium effect.

  - Page Views

    The Mann-Whitney U test showed that the number of pages viewed for each task (for Tasks 1-4) also did not differ significantly between the testing environments (lab vs NE). No significant difference was found for Page Views on All Tasks, which did not differ between lab (mdn = 23) and NE (mdn = 19.50) environments, U = 61.50, z = 1.697, p = 0.91. However, the effect size r was considered small (r = 0.3; Table 4.4).

  - Successful Task Completion

    No significance association was observed between the type of test environment and whether Task 1 was completed successfully using Fisher's exact test (p = 0.235 > 0.05). This result is also true for Task 2 (p = 1.000 > 0.05), Task 3 (p = 0.251 > 0.05) and Task

4 (p = 0.640 > 0.05), as shown in Table 4.4. The average number of successfully completed tasks per test in the lab environment (M = 2.80, SD = 1.135) was slightly higher than that of the NE environment (M = 2.20, SD = 0.894). However, a Mann-Whitney U test showed that the number of successfully completed tasks in the lab environment (mdn = 3) did not differ significantly from that of the NE environment (mdn = 2), U = 59, z = 1.95, p = 0.74 (see Table 4.5). Yet, the effect size r was considered small, r = 0.4.

- **Perceived usability**

The data collected on perceived usability for this exploratory study were subjective ratings for task difficulty and usability of the website and the number of usability issues.

- Perceived difficulty of the task

Table 4.5 shows the mean and SD values of task difficulty ratings for usability testing tasks. The Mann-Whitney U test showed no significant difference in the ratings given for task difficulty between the two environments. For Task 1, ratings given to tasking difficulty in lab environment (Mdn = 1.40) did not differ significantly from those in NE environment (Mdn = 1.00), U = 102, z = 0.109, p = .948, r =-0.3, and the effect size r is considered small. Table 4.5 shows the statistics for Tasks 2-4.

- Perceived usability of the website

In terms of overall website usability, the Mann-Whitney U test showed that no significant difference existed between the two environments. For Task 1, the ratings given to the overall website usability in the lab environment (Mdn = 2) did not differ significantly from those in the NE environment (Mdn = 2), U = 91.500, z = -0.397, p = 0.713, r = -0.3, and the effect size r was small. Table 4.6 shows the statistics for Tasks 2-4.

- Number of usability issues

Fisher's exact test showed no significant association between the usability testing environments or whether the participants reported usability issues in the entire test (p = 1.00). The Mann-Whitney U test showed that the number of problems identified in lab envrionemnt (mdn = 3) did not differ significantly from that of the NE environment (mdn = 3), U = 107.5, z = 0.338, p = .746 and r = 0.7. However, the effect size r was considered medium. Table 4.7 shows the statistics for Tasks 2-4.

Table 4. 3. Descriptive Data F and Statistical Test Results for Usability Testing Outcomes (Performance: Time Measurements)

Descriptive data (Mean: SD)	**1st Task**		**2nd Task**		**3rd Task**		**4th Task**		**All tasks**	
	Lab	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**
	(89.70: 30.76)	(120.50: 53.44)	(96.40: 47.53)	(107.73: 48.18)	(222.90: 123.87)	(232.10: 146.1)	(90.67: 21.24)	(107: 47.1)	(517.60: 140.21)	(627.11: 245.51)
Mann-Whitney U test	(U = 126, p = 0.164, r = 0.3)		(U = 126, p = 0.164, r = 0.26)		(U = 113, p = 0.588, r = 0.11)		(U = 141, p = 0.075, r = 0.52)		U = 149, z = 3.5, p = 0.432, r = 0.7	

Table 4. 4. Descriptive Data and Statistical Test Results for Usability Testing Outcomes (Performance: Page Views)

Descriptive data (Mean: SD)	**1st Task**		**2nd Task**		**3rd Task**		**4th Task**		**All tasks**	
	Lab	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**
	(3.50: 0.926)	(3.83: 1.724)	(4.90: 1.370)	(4.17: 0.514)	(7.70: 3.974)	(5.40: 3.548)	(5.20: 1.989)	4.50: 1.762	(23.10: 5.859)	(19.15: 6.072)
Mann-Whitney U test	(U = 74, p = 0.935, r = 0.2)		(U = 63, p = 0.248, r = - 0.3)		(U = 62.500, p = 0.138, r = - 0.3)		(U = 86.500, p = 0.559, r = - 0.1)		U = 61.50, z = 1.697, p = 0.91	

Table 4. 5: Successfully Completed Tasks in Each Environment and Fisher Exact Test Results

	1st Task		**2nd Task**		**3rd Task**		**4th Task**	
	Lab	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**
Percentage within testing environment group, number	80.0%	20.0%	90.0%	85.0%	20%, 2	5.0%, 1	90%, 9	80%, 16
Percentage within tasks completed successfully	44.4%	55.6%	34.6%	65.4%	66.7%	33.3%	36.0%	64.0%
Fisher Exact Test	p = 0.235		p = 1.000		p = 0.251		p = 0.640	

Table 4.6. Descriptive Data and Statistical Test Results for Task Difficulty Ratings

Rating on task difficulty	**(Mean: SD), Median**		**Mann-Whitney U test**
	Lab	**NE**	
1st Task	(1.4: 0.699)	(1.50: 0.889)	U = 102, z = 0.109, *p* = 0.948
2nd Task	(2.20: 1.619)	(1.85: 1.226)	U = 95, z = -0.244, *p* = 0.846
3rd Task	(3.90: 1.792)	(3.95: 1.508)	U = 85, z = -0.491, *p* = 0.668
4th Task	(1.30: 0.675)	(1.26: 0.452)	U = 98.500, z = 0.216, *p* = 0.875

Table 4.7. Descriptive Data and Statistical Test Results for Usability Testing Outcomes (Perceived Usability: Usability Ratings)

Descriptive data	**1st Task**		**2nd Task**		**3rd Task**		**4th Task**	
	Lab	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**
(Mean: SD), median	(2.20: 1.229), 2	(1. 0.999), 2	(3: 1.317), 3	(2. 1.071), 3	(3. 0.966), 5	(4: 1.155), 4	(1. 000), 1	(1. 0.315), 1
Mann-Whitney U test	U = 91.500, z = -0.397, *p* = 0.713, r = -0.3		U = 105, z = 0.231, *p* = 0.846, r = -0.3		U = 67, z = -1.377, *p* = 0.211, r = -0.3		U = 86, z = -0.691, *p* = 0.701, r = -0.3	

Table 4.8. Descriptive Data and Statistical Test Results for Usability Testing Outcomes (Perceived Usability: Number of Usability Issues)

Descriptive data	**1st Task**		**2nd Task**		**3rd Task**		**4th Task**		**All tasks**	
	Lab	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**
(Mean: SD), median	(0.8: 1.14)	(0.6: 0.8)	(0.3: 0.9)	(0.40: 0.7)	(1.6: 1.08)	(1.70: 1.4)	(0.3: 0.5)	(0.30: 0.5)	(0.8: 1.14)	(0.6: 0.8)
Mann-Whitney U test	U = 94.00, z = -0.300, *p* = 0.812, r = -0.1		U = 117.00, z = 1.011, *p* = 0.475, r = 0.2		U = 102.00, z = 0.091, *p* = 0.948, r = 0.2		U = 100.00, z = 0.000, *p* = 1.000, r = 0.0		U = 107.500, z = 0.338, *p* = 0.746, r = 0.7	

4.2.4.3. Type of Contextual Factors

All participants were asked to report the contextual factors experienced in the testing session; however, the lab environment participants did not report these data. Nevertheless, we were able to confirm some of those details because Loop11 reported some information about the systems used, such as the browser and IP address, which were identical for all lab environment participants.

- **Distractions**

All NE participants who experienced distractions during the test claimed they were due to multitasking or interruptions. Of the participants, 64.3% indicated that they had other software applications running* during the test. However, they claimed that they were not distracted by those software applications since they did not look at them during the test. Participants were asked to rate the distractions experienced, from 1 (to a very large extent) to 5 (to a very small extent). Table 4.9 shows the ratings for the distractions experienced based on sample size $n = 9$ for multitasking and $n = 8$ for interruptions.

Table 4.9. Participants' Ratings of the Distractions Caused by Multitasking and Interruptions

	Multitasking	**Interruption**
	(Mean: SD)	
Number of distraction instances	(1.78: 1.1)	(2.13: 1.13)
Ratings	(3.67: 0.9)	(4.1: 0.9)

Figure 4.4 shows the frequency of the types of application software that caused distractions and the maximum number of distraction occurrences per test session. The types of distraction were personal email, UEA web mail, YouTube, iTunes, chat programs, UEA portal website, user's application (i.e. word processors), system popup messages notes and demos, other website pages opened in the same browser window/tab, and other website pages opened in a different browser window/tab. These findings show that distractions occurred more often on text-based messaging application software, such as webmail,

* This is most likely to be partial as most of those participants indicated that they did not switch or shuffle between the tasks, but that the other tasks (e.g. windows) were opened in the background or minimised in the taskbar.

compared to notetaking applications and web browsers.

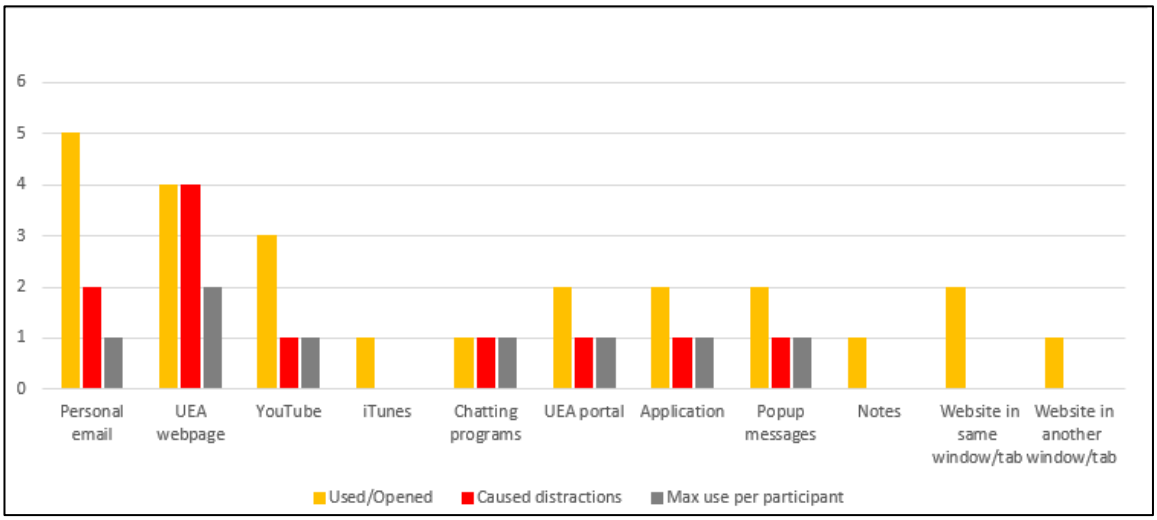


Figure 4.4: Frequency of multitasking* distractions experienced by the test participants.

Further, 45% of the NE participants were distracted by interruptions that required immediate attention, such as phone calls, text messages and responding to conversations with other(s). As shown in Figure 4.5, text messages caused the most frequent interruption during usability testing compared to phone call interruptions.

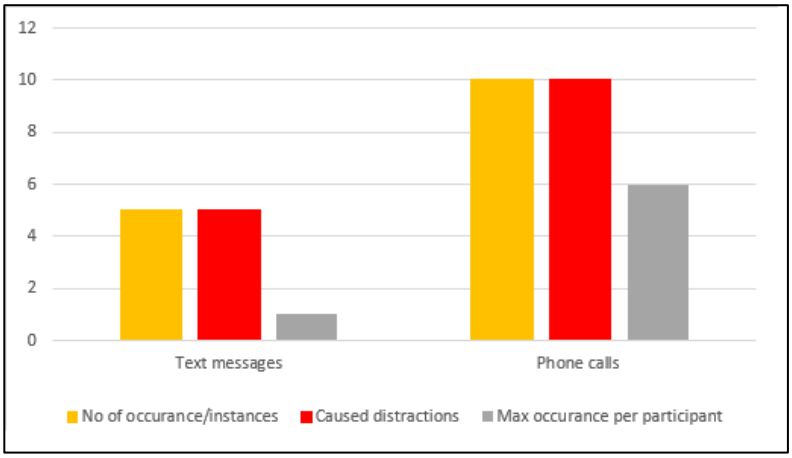


Figure 4.5: Frequency of interruptions experienced by the test participants.

The Time per Test variable summates each participant's scores for Time per Question and Time for All Tasks. Table 4.10 shows the mean and SD values of

* This is most likely to be partial as most of those participants indicated that they did not switch or shuffle between the tasks, but that the other tasks (e.g. windows) were opened in the background or minimised in the taskbar.

the completion time variable. For all variables, the mean and SD are greater in an unrestricted environment.

Table 4. 10. Time Elapsed on Questions and Test

	Lab Environment	NE
	(Mean: SD)	
Questions	(562.71: 311.82)	(1161.23: 335.95)
Test (Total)	(1099.67: 154.06)	(1572.59: 424.6)

- **Apparatus**

As reported by NE participants, 16 participants (80%) used their own laptops and four used an Android phone, notebook, tablet or PC, respectively (20%). Sixteen (80%) test participants accessed the online usability testing web portal using Wi-Fi technology via a DSL connection, one (5%) participant used a mobile phone (3G mobile connection technology) and three (15%) used a UEA network connection from their homes and offices. In terms of browsers, 13 (65%) participants used Safari web browser, five (25%) used Internet Explorer, and the remaining two participants used Opera (5%) and Netscape browsers, respectively (5%). Figure 4.6 shows a mapping between the devices, web browsers and UEA network used in the lab environment. The devices are represented by the bars and stacked by the type of network used.

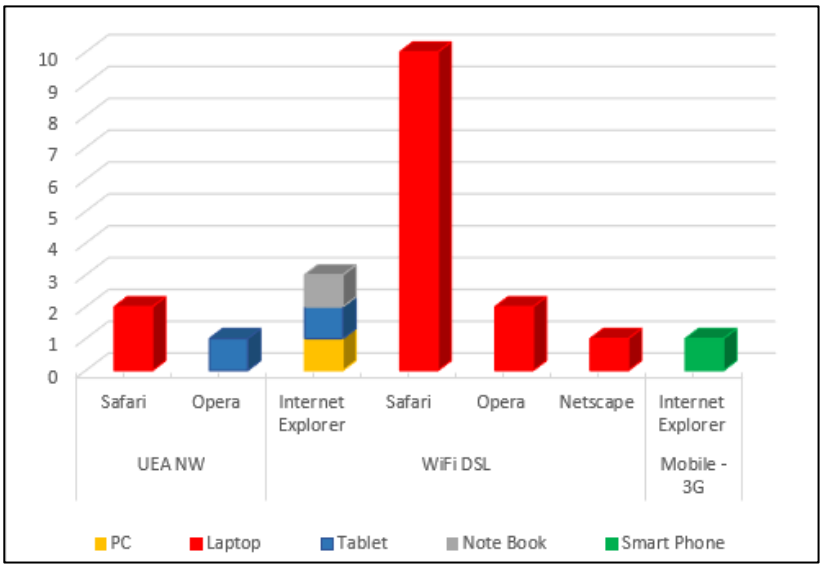


Figure 4.6: Mapping of the machines with network types and web browsers used in the NE group.

4.3 Discussion

This study explored the capability of an online usability study by adopting experimental usability testing tasks and questionnaires to collect data from remote participants in their NE and from participants in the lab environment with no observer present. The test capability was explored by determining the outcomes of a usability test in terms of performance, perceived usability measurements and data about events occurring during testing which might be considered contextual factors.

The study met its first objective by exploring the functionality of usability studies and administering the test, including the tasks, instructions and questions, within different experimental settings. In addition, for the second objective, we examined the data obtained from the participants about the interaction with the test object(s) during the testing session in the different testing environments. The third objective was to explore users' performances in different testing environment settings, which we performed by conducting appropriate statistical analyses. However, given the small sample size of the study, no statistical evidence can be given, although the findings indicate data trends.

As the study achieved its objectives, we discuss the findings with relation to RQ1:

> RQ1: What can we expect from participants of remote usability testing when they are asked to report their issues and outcomes?

Participants reported on the usability issues and distractions experienced during the online usability test. Participants in both environments who claimed that they experienced usability issues indicated how many there were and described them using Loop11 questions tool. All participants in the NE group who indicated that they experienced distraction events reported the number of occurrences and described them. Data regarding the systems used were collected from the participants in this environment using the Loop11 questions tool. The participants could choose which system specifications applied to their situation from the options given with the questions. Their subjective ratings on perceived

usability ratings and task difficulty were also collected using multiple choice questions provided by Loop11.

The participants, especially the lab group, showed a good awareness of how to answer the questions. A review and analysis of their reported data showed no conflicting data. For example, no participants mistakenly chose a small size or slow system from the options. None reported that they were distracted or were multitasking. We also asked them intentionally to specify from where they accessed the test. We also verified the participants' data to be part of the lab group from the Loop11 reports for each session, which gave the IP address of where the test was taken.

The study answered RQ2:

> RQ2: Does usability testing data performance during usability testing in the (remote) natural environment differ from that of participants in a lab environment?

Although the sample size was small and the effect sizes reported by most of the statistical analysis tests were either medium or small, the results still showed how the usability testing outcomes differ on different environments. The findings indicate that no differences were evident between the performance measurements and perceived usability in the two testing environments. This finding is a positive implication for the usability testing practice as the goal of any usability test is to allow effective performance regardless of the environment, especially when the test is unmoderated. However, further investigation is needed to examine the inconsistencies in the RAUT outcomes reported by UEM comparative studies. The fact that the outcomes of the usability testing in this study showed no significant difference stresses the importance of collecting other data related to usability testing, as discussed in Section 4.2.2, especially as some NE participants reported that they were distracted and used systems with weak performance during the testing.

This study also answered RQ3:

> RQ3: What contextual factors do remote participants experience during their usability testing session?

Some participants of NE, (64.3%), indicated that they were having other task(s) running, but those interrupted personally rated the influence of the interruption(s) experienced much negatively.

A significant difference was found in the time required to complete the test questions. Accordingly, the time for the whole test (the sum of the time for all tasks and time per question) was also significantly different between the two environments. This increased significance in time on test was likely due the differences in the time per question that was included within the time for the whole test, as Time on All Tasks was not significantly different between the test environments (Table 4.3). However, this difference might be due to different reasons. For example, the difficulty understanding the instructions in English by non-native speakers might have increased the time taken to complete the test; the comments reported by some participants indicate difficulty understanding the questions. Nevertheless, that difficulty did not influence the Time on Questions, with respect to the two environments. Notably, of the 60% native English speakers, only 13.3% were doing the testing in the lab, and the test results indicated that more time was consumed by the NE participants with a medium effect size (Table 4.11; U = 149, p = .000, r = 0.7). Further investigation is needed to corroborate this finding with a larger sample size.

Table 4.11. Time Elapsed on Questions and Test

	Lab Environment	**NE**	**Difference**
	(Mean: SD)		
Time on Questions	(562.71: 311.82)	(1161.23: 335.95)	(U = 149, p = 0.000, r = 0.7)
Time on Test	(1099.67: 154.06)	(1572.559: 424.6)	(U = 131, p =0.002, r = 0.6)

4.4 Design Limitations and Considerations

The following study design issues and lessons were identified for consideration in the following study:

- Some analysis discrepancies were experienced using Loop11 judgments to measure task success. At times, a user would believe that they had completed a task successfully, while Loop11 did not. This discrepancy required the researcher to track the clickstream of the participant to determine the actual success. To avoid this problem, participants should

be asked a question after each task about the correct answer for the task. A hint should be provided before asking whether they found the specific answer, rather than asking them to provide the answer.

- Based on the recommendations arising from the exploratory study, all the questions designed to ask about the usability issues should be placed directly after the completed task with the corresponding test object, preferably all in a single page or view.
- Task order should be randomised to avoid possible learning effects for any one task.
- A larger sample size is needed, especially to investigate the contextual factors, as it is unlikely for the whole NE group to experience distractions; a larger sample would increase the chance of having distracted participants.
- Experimental controls should be designed, adopted and applied as needed to control for any possible influence of other non-environmental contextual factors.
- As most of the study sample was familiar to Amazon.co.uk (80%), Amazon is a good choice to use to control for any perceptual influence that might arise with unfamiliar tasks.
- Only participants with a good English language level can be recruited in online usability studies conducted in English to avoid, or at least mitigate, any possible influence of language difficulties.

Chapter 5: Usability Testing Outcomes in Different Environments

5.1 Overview

The previous chapter presented the first stage of the data collection process of this research, which represented an exploratory study. We collected data using the adopted online unmoderated usability study from users reported, directly recorded data using Loop11. The data included participants' performance and subjective ratings (usability testing outcomes) and provided insights about the contextual issues involved in the testing environment.

While the findings of the previous exploratory study were promising and encouraged us to move forward with the research using the online usability study as a means of data collection, they also indicated some valuable issues to consider when designing an online usability study. In addition, the findings regarding testing outcomes (participants' performance and subjective reports) were incomplete and needed further investigation, taking into account the study's small sample size and the low-level maturity of the study's statistical design. This comparative study was conducted to avoid these negative issues and meet other objectives that will be detailed in Section 5.2.1.

The remainder of the chapter is organised as follows: Section 5.2 describes the objective of this study, presents the general design and discusses the OUUT tool used for the data collection. Section 5.2.3 presents the study analysis. Section 5.2.4 describes the study findings, and Section 5.3 presents the discussion.

5.2 Empirical Comparative Study

5.2.1 Study Objectives

In this explanatory study, we investigate the differences in the usability testing outcomes in terms of participants' performance and subjective reports. We examine what contextual factors NE participants experience and report and whether a relationship exists between the usability testing outcomes in terms of participants' performance and subjective reports and the contextual factors reported. This study answers the second, third and fourth RQs:

> RQ2: Does usability testing data performance during usability testing in the (remote) natural environment differ from that of participants in a lab environment?

RQ3: What contextual factors do remote participants experience during their usability testing session?

RQ4: How do the contextual factors influence the users' outcomes during usability testing?

To answer these RQs using OUUTs, this study seeks to meet the following objectives:

- Redesign the online unmoderated usability study to avoid the issues found in the previous exploratory study and apply the suggested design features.
- Enhance the statistical design of this comparative study to avoid or mitigate the limitations discussed regarding the previous exploratory study.
- Investigate the contextual factors reported by remote participants during their RAUT session.
- Investigate the difference in usability testing outcomes in terms of participants' performance and subjective ratings in different testing environment settings.
- Investigate the relationship between the contextual factors reported by participants and the differences in the usability testing outcomes, if any.

By redesigning, enhancing the statistical design and conducting the study, we aim to meet the above objectives and answer the RQs.

5.2.2 Study Design

To answer the aforementioned RQs, we design an online usability study that applies RAUT, which is accessible by participants in different environmental settings at the same time, as in the previous exploratory study. As depicted in Figure 5.1, there are two groups in two experimental conditions: lab and NE participants perform the usability testing tasks through an identical online unified access port. In this study, we collect data on the required measurements and design the experimental tasks, procedures, and statistical design and controls. The data collection method or means is specified. With respect to the required measurements to answer the RQs, we collect data on the testing outcomes and contextual factors. Based on the 4FFCF model, usability testing outcomes are represented in participants' performance and perceived usability (see Figure 5.2). The measurements adopted for the participants' performance in this study were Time on Tasks, Page Views, and Successful Completions. Subjective reports were collected to measure the perceived usability and subjective reports on usability issues (see Figure 5.3).

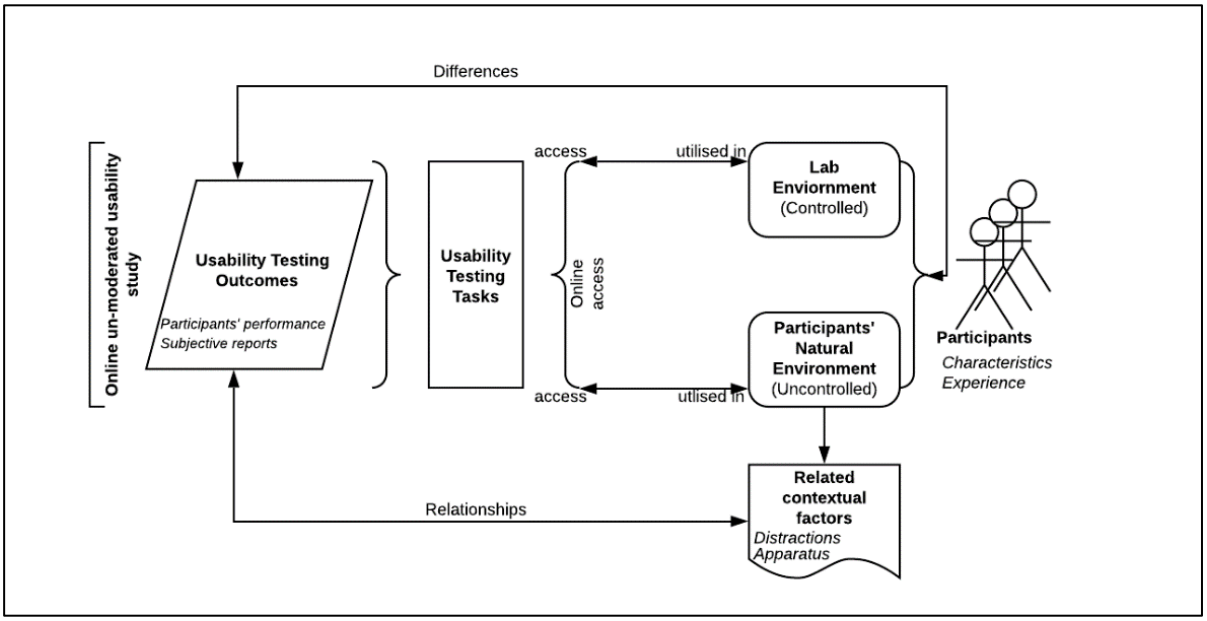


Figure 5.1. Comparative study design.

Usability testing outcomes are influenced by four main factors: testing environment, user characteristics, task scenario and system prototype. While we can experimentally control for task scenario and system prototype for both environmental settings, it is impossible to control for user characteristics in the adopted design depicted in Figure 5.1. The study design implies that a different group should be allocated to each testing environment, which means the data collected for each testing session is carried out by different participants. The dominant between-subjects experimental design of the study (Figure 5.1) necessitates the need to have participants in different groups that are as homogeneous as possible (Lazar et al., 2010).

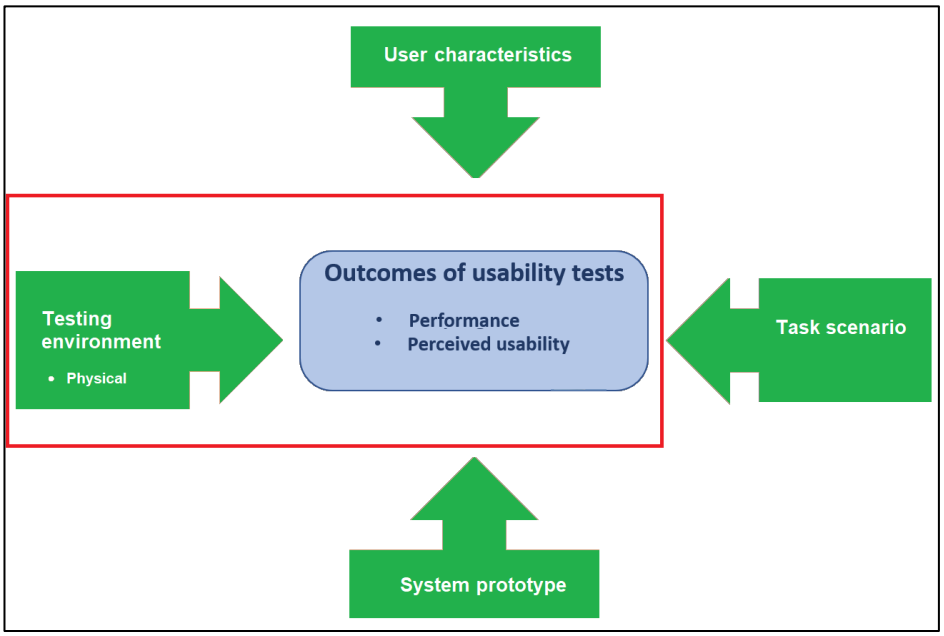


Figure 5.2: The factors to be empirically investigated and validated by the 4FFCF model in this study.

As a result, besides gathering data about the environment (what happens in the testing session), data regarding participants' characteristics are also collected (Figure 5.3). The aim is to apply experimental and statistical control techniques as is required and relevant to avoid or mitigate the influence of participants' characteristics. The following sections provide more details.

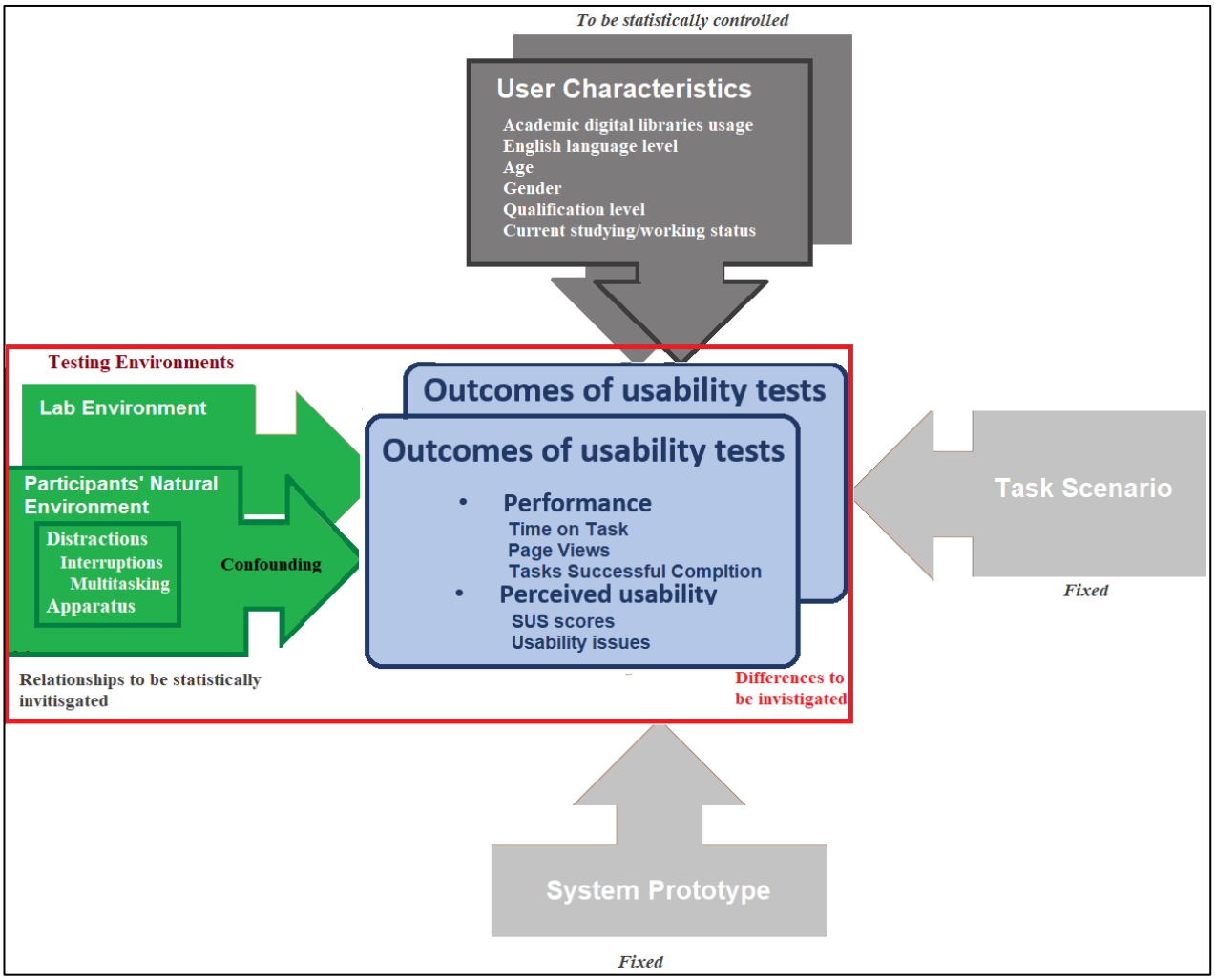


Figure 5.3. Experimental design with respect to the 4FFCF model.

To ensure the performance data are objective, the data are recorded automatically or at least partly derived from automatically recorded data, as was done in the previous exploratory study using the OUUT tool.

With regards to the perceived usability of the test object(s), the self-developed scales used in the previous study were replaced by a standard usability scale tool, which is already implemented and used in the literature. The reason behind this change is to use a much more reliable tool to collect more accurate data and to increase the generalisation of the results' comparison in the future.

Therefore, we selected the System Usability Scale (SUS) to collect participants' subjective ratings. The SUS is probably the most popular questionnaire used for measuring attitudes towards system usability (Lewis, 2006; Zviran et al., 2006). The SUS is generally applicable regardless of the technology used (technology-neutral; Brooke, 1996). The SUS consists of 10 items that alternate between positive and negative statements about usability; the odd items are designed to be positive, and the even items are negative. The response options ranged from 1 (strongly disagree) to 5 (strongly agree), as shown in Figure 5.4.

The SUS has been acknowledged as a good choice when the benefits of alternating the wording of items outweigh the potential negatives (Finstad, 2006; Bangor et al., 2008; Finstad, 2010; Lewis and Sauro, 2009).

Regarding usability issues, the questions were designed to ask the participants to report usability issues, such that they would indicate their existence, how many there were and list them. Based on the recommendations arising from the exploratory study, all the questions designed to ask about the usability issues were placed directly after the completed task with the corresponding test object, preferably in a single page or view.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The response options, arranged from the left to right, are Strongly Disagree (1) to Strongly Agree (5).

Figure 5.4 System Usability Scale (SUS).

In both environments, participants were able to indicate whether they were distracted and list any distractions. In addition, they were asked to report the type of system used. However, based on the findings of the previous exploratory study, it was expected that only NE participants would report the data for distractions (interruptions and multitasking). Gathering such data helps to study the relationship between the differences in testing outcomes and environmental factors.

Participants' characteristics were represented in data on the participants' demographics and experiences. Some demographic and experience data were collected before testing to act as experimental controls and filter out the study participants or select appropriate sampling techniques to apply. The other data were collected at the end of the study to be used in the analysis to determine whether their characteristics influence the data and, if so, adopt appropriate statistical techniques to consider or exclude that influence in the desired study analyses activities.

In line with Greifeneder (2011), the results of the exploratory study indicated a possible influence of age and prior knowledge (Vakkari, 1999) on the type of tasks to be carried out in the experiment or test. In this study context, prior knowledge of tasks would mean that a participant's academic speciality would be similar to the type of website used to perform the task. Besides the typical experience data collected in usability testing experiments, for this study, academic digital libraries' usage, experience with usability testing, English language level, age and academic speciality were collected before the test using an online screening questionnaire that was accessed by candidate participants; this can be done using any online surveying tool which supports question branching.

Other characteristics that were collected after the test are gender, qualification and current studying/working situation. All the aforementioned data are collectable using the OUUT tool described in the following section.

5.2.2.1 OUUT Tool: UsabilityTools

UsabilityTools is an online unmoderated usability tool (www.usabilitytools.com) used to design, administer and launch online usability studies. Both UsabilityTools and Loop11, which was used in the previous study, are affordable tools that enable the design of tasks and questions for usability testing. Both tools allow for automatically recording data on time spent on tasks and time spent on questions. Both record the visited page URLs and

neither require the participants to install additional software on the machine. Additionally, neither tool requires the participants to switch between windows to view the questions and the testing objects, as they are fully functional within the test window. However, both tools restrict transfer from the web testing page (task page) to the next page unless one of the two buttons 'Success' or 'Give Up/Abandon' are pressed. Both also show the task at the top of the window of the object being tested and enable the designer to provide the successful URLs in the design stage of the test before launching the test; these track the usage accordingly and indicate whether a participant has succeeded.

We opted to use UsabilityTools for this study, rather than Loop11, because it provides a platform for designing different forms of unmoderated usability testing from a conversion suite, a user experience (UX) suite or the voice of customers. The UX suite allows implementation of any one of the UX tools, such as survey page(s), web testing page(s) (task scenario page(s)), and other pages for other testing types (e.g. card sorting or click testing). The UX also enables more than one of these tools to be used in the same test/study (UsabilityTools, 2016). This capability was a highly valuable design criterion for this study because one of the limitations found in the previous exploratory study design was the inability to ask more than one question after each task. When several questions had to be presented on multiple pages, the chance of forgetting what happened in the past task would increase. However, UsabilityTools UX suite allows an entire page of survey to be designed with any number of distinct types of questions. This criterion is useful for asking multiple questions just after a task's performance (e.g. questions regarding test experience and the type and number of usability issues).

UsabilityTools also provided more capabilities for writing and presenting descriptive instructions and provided a much larger space to enter text. This criterion is also valuable, especially with the absence of the testing moderator. UsabilityTools allowed for conditional logical branching, which was not available with Loop11 at the time. This feature assists in designing screening questions and other questions that require branching. UsabilityTools also enables the designer to locate the questions before or after the tasks as desired. As a result, instructions were located before the tasks, and questions asking about the experience with the task were asked after each task (Figure 5.5).

- **Pilot test 1**

Because there was a need to test how UsabilityTools works and functions in a real-time testing situation, a small pilot test was conducted with seven volunteers to verify that UsabilityTools was technically acceptable and functional with multiple browsers and devices. UsabilityTools was found to be technically and functionally acceptable; however, one limitation found was in its inability to exclude further access to the same IP address and checking entered IDs, which Loop11 provided. As a result, the experimental control was manipulated by the researcher using the screening process described in Section 5.2.2.5.

5.2.2.2 Experimental Design and Tasks

The between-subjects variable refers to the two environmental settings: lab and NE. The within-subjects variable was the four tasks: Task 1, Task 2, Task 3 and Task 4.

As in the previous exploratory study, three digital libraries – the Universal Digital Library (UDL), Perseus Digital Library, and arXive Digital Library – were used to perform the tasks on, along with Amazon.co.uk, which served as a control website. In addition, a task on the Digital Public Library of America (DPLA) was designed to train and familiarise users with the test requirements (see Table 5.1).

Table 5.1: Test Objects Used in The Study

Task Type	Target Website	URL	Specialty(ies)	Provider
Training Task	Digital Public Library of America (DPLA)	http://dp.la/	General	Harvard University
Task A	The Universal Digital Library (UDL)	http://www.ulib.org/index.html	General	Carnegie Mellon University
Task B	Perseus Digital Library	http://www.perseus.tufts.edu/hopper/	History, literature & culture of the Greco-Roman world	Tufts University
Task C	arXiv Digital Library	http://arxiv.org/	Mathematics, physics, computer science, quantitative biology & statistics	Cornell University
Task D	Amazon UK	http://www.amazon.co.uk/	Sales	***Founder:*** Jeff Bezos

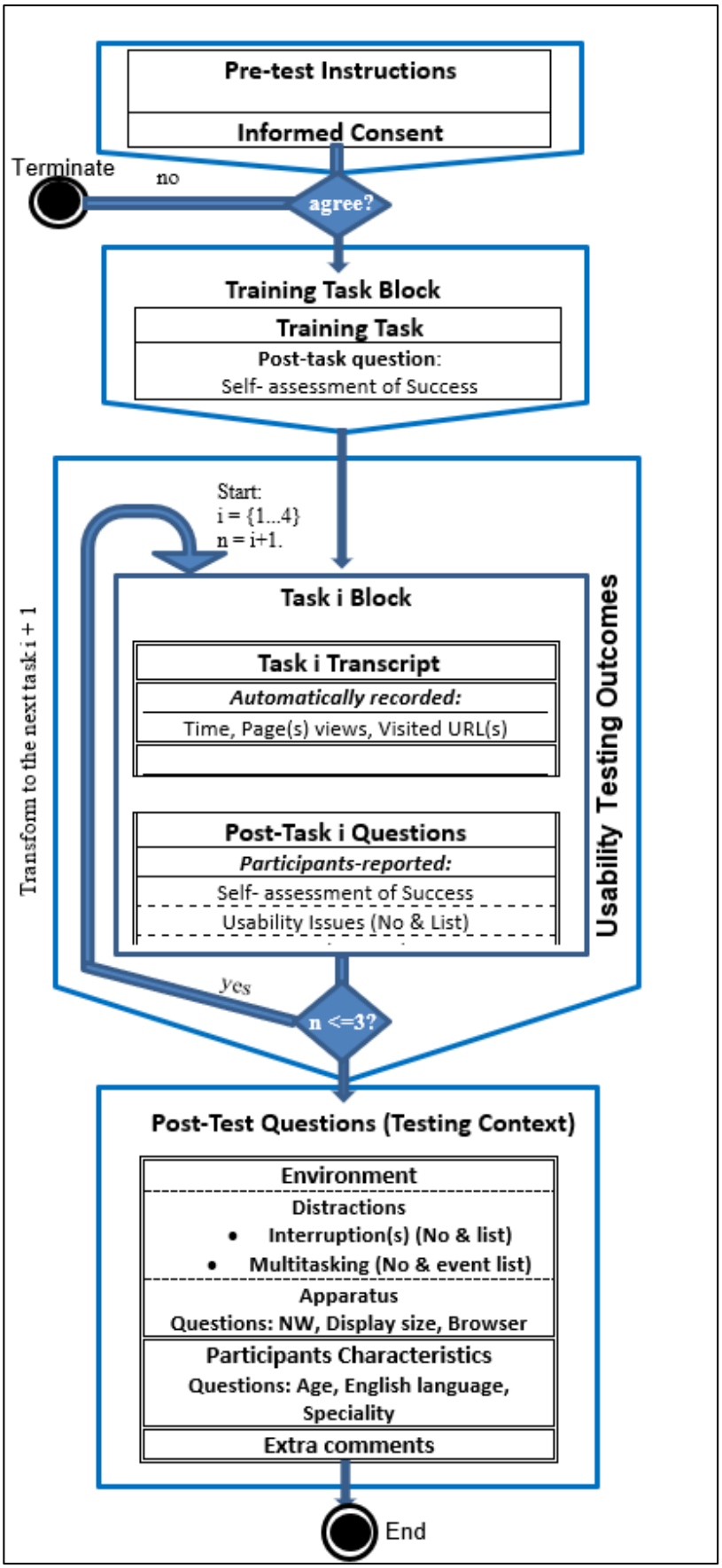


Figure 5.5: The navigation of the data collection process through UsabilityTools.

Participants were asked to complete one predefined task for each digital library website. The nature of an experimental comparison necessitates having predefined tasks to allow for comparisons of performances between the different environment settings. A transcript of the tasks can be found in Appendix CH5.1.

In addition, to obtain more generalised results and to determine whether participants' performances would differ with regard to different tasks of different complexity levels among the different environmental settings, we set multiple tasks with different complexity levels according to the elements specified by Campbell (1988).

For example, the task for the Perseus Digital Library was perceived as low complexity because one path could be followed to reach the target. The task with UDL was perceived as medium complexity because there was uncertainty or ambiguity about the path needed to reach the target. The arXive task was perceived to be complex, as there were multiple potential paths to reach the target. The task also seemed to have multiple targets, but only one target was correct, and there was potential uncertainty or ambiguity about some paths.

While complex tasks are difficult, tasks can be difficult (i.e. require high effort) without being complex (Campbell, 1988, p. 45). The perception of task difficulty relates to the psychological state of the individuals performing the task (Campbell, 1988). In addition, in some cases, individuals require advanced skills to navigate poorly designed websites, and some might lack the background knowledge needed to understand some tasks. Thus, task complexity might relate to the nature of the task, the individual's attitude or both.

- **Pilot test 2: Tasks design review**

To decide the complexity of each task and based on the information discussed earlier in Section (5.2.2.2), we conducted a review for the design (Tasks Design Review 1) with 16 volunteer participants (62% female) aged between 22 and 30 years (Mean = 25.81; SD = 2.71).

Participants were required to rate the tasks before and after performing them. A pre- and post-experimental design allowed for identifying whether the difficulty ratings assigned to a task were based on the individual's attitude towards the task (participants' ratings to the task complexity before the performance) and after the performance of the task (due to the complexity elements inherited within the tasks). If the individual ratings were consistent

before and after task performance, we argue that this should indicate that the ratings reflected the complexity of the task rather than due to the poor usability of the website.

The participants were not timed while performing the tasks, and no usability testing method was used. Instead of being asked to provide answers for tasks, the participants were asked to stop working on the task when they believed they had found the answer or would not be able to find it. The participants were recruited from the same population as the sample of participants for the formal empirical study.

Ratings were done using the Single Ease Question (SEQ), which was chosen because it is considered reliable, sensitive and valid. SEQs meet the four characteristics of a good questionnaire: (1) short, (2) easy to respond, (3) easy to administer and (4) easy to score. SEQs can be administered on paper, electronically or even verbally (Sauro, 2010).

Figure 5.6: Single Ease Question (SEQ) (adapted from Sauro, 2010).

Comparisons of the SEQ with other questionnaires (e.g. UME* and SMEQ†) have shown that the SEQ performs very well (Sauro and Dumas, 2009; Sauro, 2010). The SEQ used in this design review was in the form of a paper questionnaire and respondents answered on a seven-point scale, ranging from 1 (very difficult) to 7 (very easy). Figure 5.6 shows an example of an SEQ question.

The Wilcoxon Signed Rank statistical test indicated no statistical significance between the ratings before and after the performance of any task. In addition, Kenall's Tau b‡ showed a

* Usability Magnitude Estimation

† Subjective Mental Effort Question

‡ Kenall's Tau b is similar to Spearman's correlation as '[t]his test is still used for cases where at least one of the variables include non-parametric data. The main difference is that Kenall's Tau b should be used if there are too many tied ranks. How many is too many? There is no golden rule' (Mayers, 2013, p.121).

significant concordance between ratings in both pre- and post-task performance conditions for each task. Table 5.2 presents the results of the pilot study.

Table 5.2: Statistics for The Task Design Review1

	Median Value of the Pre-performance Ratings	Median Value of Post-performance	Statistics for Wilcoxon Signed Rank test	*P*-value of the Kenall's Tau b
UDL	5	4	Z = -1.21, *P* = .227, r = 0.20	0.798
Perseus	7	7	Z = -1.41, *p* = .157, r = 0.24	0.537
arXive	2	1.5	Z = -1.00, *p* = .317, r = 0.17	0.882

The p-value of significance is at 0.05

The median values presented in Table 5.2 show that participants' ratings are consistent before and after their performance. The overall results* also show that the level of complexity ratings given for each task vary between low, medium, and high (refer to the median value for the ratings before and after the performance). The variation in task complexity enables the study to investigate whether task difficulty has different influences in different environments. Appendix A.CH5 presents the transcripts for the tasks.

5.2.2.3 Experimental Conditions

As mentioned, there were two experimental conditions: lab and NE environments. Neither experimental condition had an observer or 'test monitor' (no direct/physical observations) or passive observation (video/audio recordings; Figure 5.7).

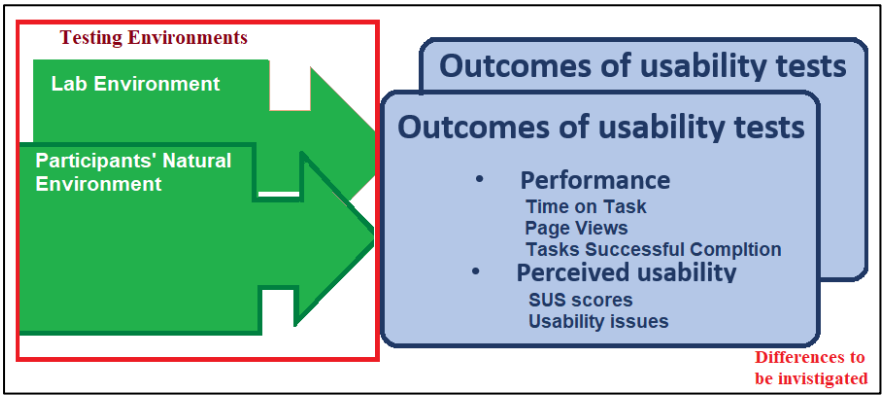


Figure 5.7: Experimental conditions outlined by the red box.

* We did not relate to the previous ratings for tasks difficulty that were given in the previous study because two of the test objects (the digital libraries' websites) were unable to function with the newly used tool (UsabilityTools) in this study. These objects were Jstor (http://www.jstor.org/), which was used for a training task, and CiteSeerX (http://citeseerx.ist.psu.edu/index). Thus, we designed different tasks which necessitate a new design review.

NEs were considered to be any environment in which the test participants could access the online usability study. No restrictions were placed on the type of computing device or smartphone, the browser, and the Internet access or network the participants could use to access the usability study and perform the test (Figure 5.8(a)). However, for the lab environment, participants were restricted to using only the assigned system (Figure 5.8(b)). Table 5.3 presents the details of the system used in the lab environment.

5.2.2.4 Study Advertisements

Several study advertisements were designed and published using several means, including classical means, such as flyers and posters, and emails and social media, such as Facebook and Twitter.

The email content included the study's purpose, importance, guarantee of data confidentiality, consent information, test duration, incentive amount, method of receiving the incentive and the researcher's email address to contact the researcher if interested in participating.

Participants were told that the aim of the study was to improve the usability of digital libraries because participants were not supposed to know that there were different environmental settings. The Facebook post content was identical to the posters and the emails. The content of the A3 flyers was a summary of the information in the recruitment emails and on the posters.

Twitter was used to broadcast a very brief text, including the researcher's email. The email used the official UEA webmail system using UEA mailing lists from multiple schools. A4 posters were placed on multiple UEA bulletin boards and contained identical content to the recruitment emails.

A3 flyers were distributed throughout the UEA campus, library and UEA school hubs. In addition, the social media accounts related to UEA were targeted using the UEA network. Appendix CH5.2 shows an example of the advertisement materials.

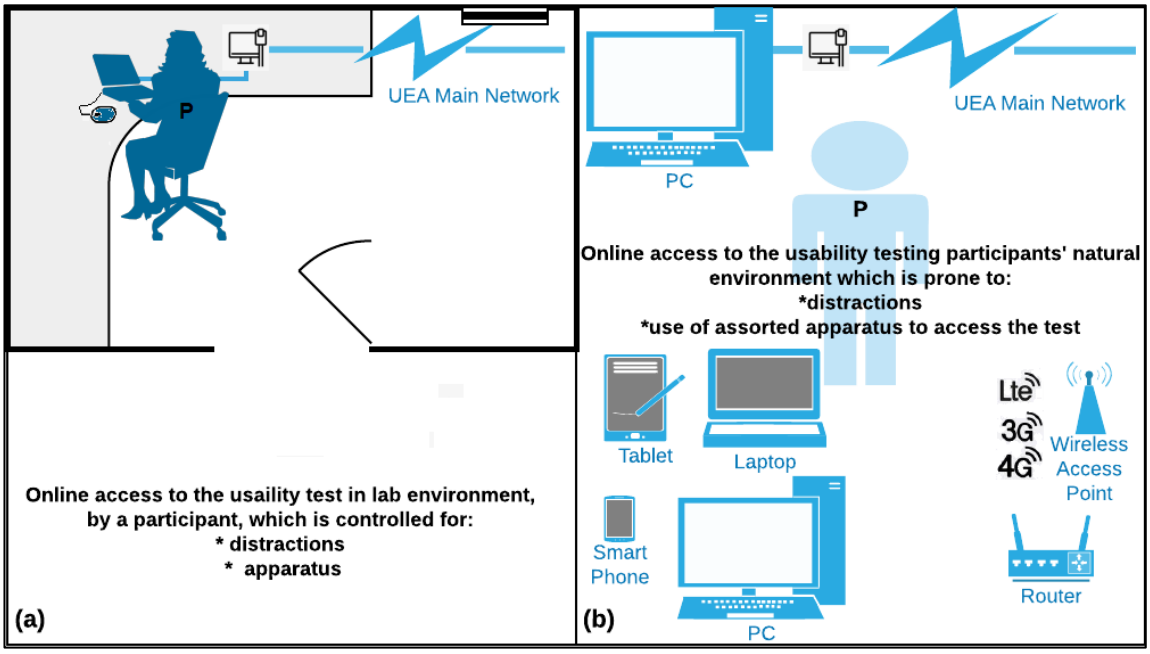


Figure 5.8: Setup of each testing environment. (a) lab setting, (b) model of NE settings. (P = Participant)

Table 5.3: System Specifications Used by Lab Environment Participants

	Description
Machine	Laptop, Intel® Core™ i5-232OM CPU @ 260Hz
Operating System	Windows 7, 64-bit
Browser	Google Chrome Version 49.0.2623.112 m (64-bit)
Internet Connection	UEA Main Network, Fast and Reliable
Additional Requirements	Wireless Mouse, Logitech

5.2.2.5 Experimental Controls

- **Information disclosure control**

The study was advertised after receiving ethical approval. To eliminate the possible bias that might affect their performance, no information or instructions regarding reporting distractions or the types of systems used were given in advance of the experimental tasks. In the test advertisement materials, participants were only told that they could indicate their interest in participating for a two-week period.

To decrease the probability of recruiting profit-seeking participants, participants were informed in the advertising materials that the vouchers would be subject to availability and would be delivered by email. Participants were given the options of voluntary participation

or a £7 Amazon.co.uk voucher incentive; they were allowed to choose whether they wanted to participate as volunteers or wished to receive incentives to limit the possibility of sample errors due to over-motivated or profit-seeking participants.

- **Participants with certain criteria**

Based on the previous exploratory study suggestions, this study only accepted participants based on the following criteria:

- ✓ Students from the UEA who responded using their university email.
- ✓ Students who had used any digital library website at least once a year prior to enrolment. This criterion was added to control for any negative performance associated with a lack of knowledge or experience with digital library websites.
- ✓ Students have not participated in or had any prior experience with usability testing.
- ✓ Only native English speakers and participants from a non-English speaking background who considered themselves either 'fluent' or 'moderately fluent' in English. A sufficient proficiency in English was required for reading and understanding the tasks, websites and questions, which were all in English.

- **Homogenous groups**

Sufficient inclusion of both age groups and specialities in the sample was ensured as acknowledged before (Section 5.2.2.2). This was attainable by using a randomised blocks design, which helps to reduce noise or variance in the data. A randomised block design based on age and speciality was applied to the sampling process because these criteria were requested in the screening questions before formally enrolling for the test (see Appendix CH4). The sample was divided into relatively homogeneous subgroups, or blocks, based on age groups (18-24 and 25-34 years) and academic specialities (text-oriented and mathematically oriented). The results obtained eight blocks of 12 participants, which were then randomly allocated to the lab or NE group (Table 5.4).

This sampling technique ensured that the experimental design was implemented within each block or homogeneous subgroup. As such, the variability within each block was less than the variability of the entire sample, and each estimate of the treatment effect within a block was more efficient than estimates across the entire sample. When the more efficient estimates were pooled across blocks, an overall more efficient estimate was obtained than without blocking (Leedy and Ormrod (2005).

Table 5.4 Randomised Blocks Sampling

Age group		**Academic Specialty**		**Resulting Blocks**	**Randomly allocated to**
(18-24)	N: 48	Text-oriented	N: 24	N:12	Lab
				N:12	NE
		Mathematically oriented	N: 24	N:12	Lab
				N:12	NE
(25-34)	N: 48	Text-oriented	N: 24	N:12	Lab
				N:12	NE
		Mathematically oriented	N: 24	N:12	Lab
				N:12	NE

For the lab environment, participants were instructed in the email to head to the experiment room, which was a small, quiet room reserved in the UEA Computing Science School. No distractions were allowed and all participants who carried out the experimental test in this environment used the same systems – the computing device and online communication means and technologies (Table 5.3). Only the Google Chrome browser was used and was pinned to the taskbar. The lab participants were instructed verbally before entering the lab testing room that distractions and multitasking were not permitted while taking the test. This rule was also presented on an instructional poster posted in front of the participants in the testing room. Participants were verbally instructed by the experimenter (the researcher) to use the machine provided on the desk in the reserved room to access the experimental usability test page through the web-portal which was already prepared and open in the browser's window. The machine was standardised so that only the web browser used in the experiment was available and the desktop had no visible files or programs that could be used.

For the NE experimental condition, participants were instructed in the email to take the experimental test at a time that suited them in one continuous session within the two-month period when the online page for the experimental test would be open. A link to the web portal was given to the participants who met the screening criteria and were randomly allocated to the NE environment. No instructions were given regarding contextual factors (e.g. distractions and the type of systems might be used) because they might affect the ability to capture the real situation and the context of the test participants, which would ultimately affect the validity of the experimental comparison. As such, the participants were not informed that distractions were not permitted or that they were restricted to a specific type of system. Additionally, they were not informed that distractions were

permitted or that the use of any type of system was permitted. Rather, the instructions regarding these issues were undisclosed to avoid the possibility of bias in the experimental results.

- **Access control**

The participants enrolled in both groups (Lab and NE) were informed that they would only be able to access the test if they provided the enrolment ID given to them in the participation approval email, which was sent before the test (see Appendix A.5.1, Figure A.4). The enrolment ID was formulated to have 12 digits. The first digit reflected the index of the first block (the age group), which was either 1 or 2, and the second digit reflected the index of the second block (the academic speciality), which was also either 1 or 2. The following two digits were the participants' IDs, and the last eight digits reflected the encrypted forms of the eight digits of the UEA User ID* (which was the first eight digits of the UEA email address)†. Including the first eight digits of the UEA email address guaranteed uniqueness, as no student or member of UEA had the same first eight digits in their UEA email. The UEA digit encryption guaranteed that participants could not have inappropriate use of the assigned enrolment ID.

Encryption was necessary so that participants could not infer that these digits referred to the UEA ID digits‡. For alphabetical digits, simple encryption was used (e.g. A became Z, and B became Y). However, numerical digits were encrypted alphabetically (e.g. 1 became A, 2 became B and so on) and not simply by reversing them (e.g. 1 became 9, and 2 became 8)§. Participants were asked for their enrolment ID again at the beginning of their test session, through UsabilityTools (Appendix CH5). UsabilityTools kept a record of the enrolment IDs so we could relate some of the screening data with the testing outcomes (see Appendix CH5).

* UEA user ID is not the student ID. The student's UEA ID can be found by their UEA email address, as it constitutes the first part of the email address (the part that precedes the @ mark).

† The file that included this information has been encrypted and saved in external storage.

‡ For example, participant X begins chatting with his friend about recently being recruited for a usability experiment and that he was given an enrolment ID, which includes his UEA User ID. If the participant selected voluntary participation, he is likely highly motivated to participate and unlikely to expose the enrolment ID. If the participant chose the incentive, then he is also unlikely to expose the enrolment ID to his friend. Inappropriate use might occur if the participant informs the friend that the first part of the UEA email was included. This might cause the participant's friend to attempt to login using the other student's UEA User ID.

§ The intention behind this is that we do not want to apply reversing the digits to the same data types for both the numerical and alphabetical part of the UEA ID (reversing numbers to other numbers and letters to other letters) so we ended up with an encrypted UEA ID that might resemble a real current ID for an unknown student.

- **Learning control**

The task order might have a significant impact on the results, as participants usually learn the system as they gain experience, known as the 'learning effect' (Tullis and Albert, 2013; Albert et al., 2009). Randomising the order of the tasks cancels out potential errors introduced by differences in tasks (Lazar et al., 2010). Lazar et al. (2010) argued that regardless of the experimental design adopted, it is important to counterbalance the orders of the tasks.

UsabilityTools does not provide the ability to randomise the tasks, unlike expensive tools (e.g. UserZoom). However, as UsabilityTools' price plan is pay as you go, it was possible to design eight versions of the usability study, four versions for the online usability study to be administered in the lab, and four versions to be administered in users' NEs. Each version had a specific task order (see Table 5.5). Versions 5-8 are repetitions and assigned for online usability to be administered in the NE. By creating different versions, we ensured that equal divisions of the whole sample were performing tasks in a distinct sequence for every experimental setting.

That is, each block of 12 participants of a specific group (refer to Table 5.3) was then categorised into four groups of three participants, and each group was assigned to one of the four versions of the online usability studies.

Table 5.5. An Example of Random Allocation of the Experimental Tasks for an Experimental Condition

Version	**Task A**	**Task B**	**Task C**	**Task D**
Online Usability Study Version 1	Perseus	UDL	Amazon	ArXive
Online Usability Study Version 2	UDL	Amazon	ArXive	Perseus
Online Usability Study Version 3	Amazon	ArXive	Perseus	UDL
Online Usability Study Version 4	ArXive	Perseus	UDL	Amazon

- **Data anomaly control**

Time on task takes longer if technical issues occur. This extra time for the task performance time arguably does not reflect a genuine contextual factor related to the difference between the Lab and NE conditions. In addition, if participants have previous experience with the test object, i.e. the website, used for the underlying task, the performance might be influenced,

most likely positively, as the participant will be familiar with the website layout and functionality. However, these aforementioned issues could not be addressed until the task was completed. That is, participants were asked after completing each task block whether they had previously used that website. They were also asked to report any technical issues they faced while completing the task. The answers to these two questions thus enabled any corresponding data scores from related statistical analyses to be adjusted.

- **Incentives delivery control**

The incentive amount was the same for both environmental settings. The incentives were delivered via email for two reasons. First, email is the best way to deliver the incentives to the online participants, especially those who performed the test in their NEs. Second, email delivery ensures that only those who received participation emails received the incentives after participation.

To decrease the probability of recruiting profit-seeking participants, participants were informed in the advertising materials that the vouchers would be subject to availability and would be delivered by email. Just before the end of the experiment, each participant was asked to provide an email address for the delivery of the incentive or to skip the email question if they wanted to opt for voluntary participation. The email address for the incentives was immediately separated from the dataset and stored as an encrypted file on an external hard disk.

5.2.2.6 Ethical Clearance

The data collection design shown in Figure 5.8 using UsabilityTools and the advertisement design were ethically approved before commencing the experimental procedures. Before seeking ethical approval, several pilot tests and redesigns were made (e.g. the previously mentioned pilot tests 1 and 2). Once the experiment was fully designed, all the documentation, including the required participant reassurances, screenshots of the study design materials and informed consents were submitted to the Ethical Approval Committee in Computing Science School in UEA. A few adjustments were made to the data collection methods and advertisements after obtaining the final approval for the designs (see Appendix CH5).

5.2.2.7 Experimental Protocol

After receiving ethical clearance, the experimental protocol was started. As shown in Figure 5.9, most of the experimental control was applied before starting the experimental procedures.

The students expressed their interest in participating in the study via the email address provided in the study's advertisements. Then, the online experimental controls were applied (Figure 5.9).

The participants received a screening questionnaire that was designed using UsabilityTools (Appendix A.4). After screening and sampling the participants, the selected participants received an email confirming their participation along with their enrolment ID. The selected participants' data were associated with their assigned enrolment ID and saved in a spreadsheet.

The test period lasted two months, during which time prescheduled appointments were offered to participants assigned to the lab environments. Scheduling was carried out so that each participant was assigned one hour, based on the pilot studies, for the lab room in a time agreed between the researcher and the participant. Participants who were assigned to the NEs were informed in the participation approval email that they could complete the test once, in one continuous session, within two months.

The enrolment ID was verified twice. The first time was via the web portal to assign each participant to the appropriate online usability study based on the tasks sequence pattern. After the participant accessed the desired online usability study, their enrolment ID was obtained for the second time by UsabilityTools, which saved it to enable aggregating usability testing data and screening data later.

UsabilityTools guided the participants through the experimental test. The test started with a welcome page where participants were instructed to give their online consent before starting the test session to confirm their willingness to participate. The welcome page presented an overview of the purpose and nature of the experimental test and other information about the test. The participant was only allowed to proceed with the test session if they agreed to give their consent.

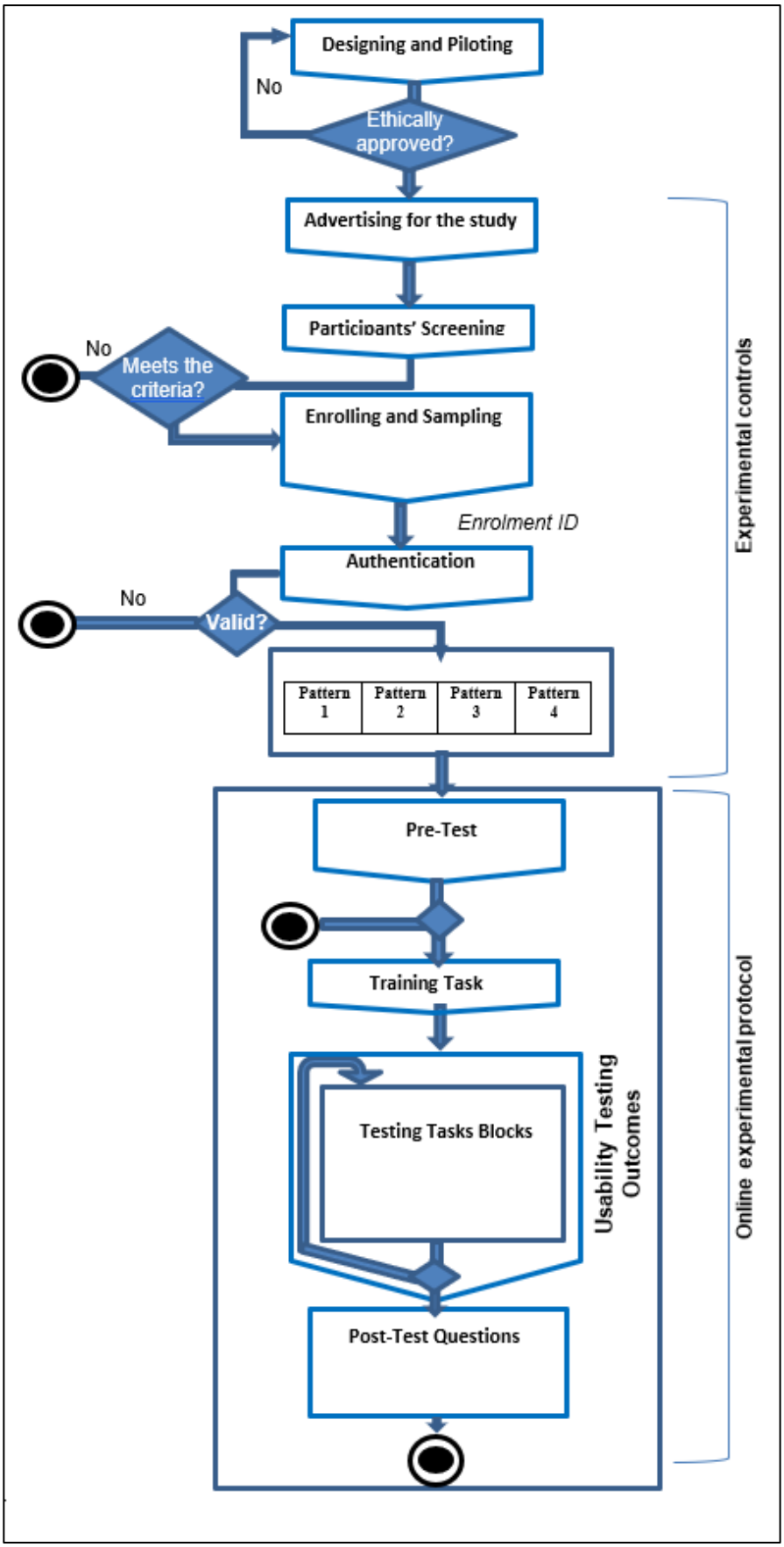


Figure 5.9. Online experimental controls and protocol.

Then, using UsabilityTools, the participants were instructed to perform the tasks and answer the questions honestly. They were informed that their answers would not affect their participation incentives to reduce the possibility of social desirability responses.

No observer was physically present in either experimental condition (Lab and NE). Participants were guided by UsabilityTools to carry out the training task and they were informed that they did not need to provide answers to the training task. Instead, they indicated whether they thought they had found the answer, which justified not using the time recorded for the training task in the analysis[*]. Then, they were asked to perform the actual timed tasks (UDL, Perseus, arXiv and Amazon), answer self-assessment questions relating to their success after each task and answer the control questions (to indicate whether they had previous experience with the website or faced technical issues during task performance).

After completing all the tasks, the participants were asked about the contextual factors (interruptions and multitasking instances) and their characteristics (demographics and experience).

Last, the participants had the option to comment and to provide their email address to receive the incentive (Appendix A.CH5). They were asked to allow a maximum of 48 hours for incentive delivery and were advised to contact the researcher if they had not received anything in that time. Finally, participants were presented with the final page, where they would realise they had finished their experimental test and where contact information was given (see Appendix A.CH5).

5.2.3 Study Analysis

Overall, the analysis activities for this study were carried out in three stages. The first stage involved preparing the data for use in the analysis procedures. Then, the data were explored to obtain insights about how the data are distributed. Last, the analysis procedures were carried out to answer the RQs.

[*] The time was included within the 'time spent on test' to reflect the test experiences of all participants.

5.2.3.1 Data Preparation

As data were collected through UsabilityTools, the spreadsheets for the different study versions (based on the tasks sequence pattern) were named according to the experimental conditions and version. SPSS 22.0.0.0 data statistics were used to read the data, perform the required statistical analyses and code the data. Then, a quality check of the data was carried out. Extreme value and data outliers were investigated, and the necessary adjustments applied. Reliability checks were applied to the SUS scale. As shown in Table 5.6, the SUS scale has good internal consistency for every task with respect to each experimental condition, and for the whole sample, as all the values were above 0.7 (DeVellis, 2012).

Table 5.6. Cronbach's Alpha Coefficient Values for SUS Scores for Each Task in Each Environment and for the Whole Sample

Experimental conditions	**Task A** ***Perseus***	**Task B** ***UDL***	**Task C** **a*rXiv***	**Task D** ***Amazon***
Online (Lab)	0.800	0.813	0.880	0.909
Online (NE)	0.795	0. 806	0.868	0.909
Whole sample	0.792	0.823	0.870	0.906

After preparing the data, the data were checked to see if they had a normal distribution. If the data were found to not be normally distributed, data transformation techniques were used, if applicable, to transform the data. Then, appropriate statistical analysis tests were selected based on the data nature and the type of the RQ to be answered.

5.2.3.2 Data Exploring

Ninety-six participants were recruited for this study (48 participants in each experimental condition). The distribution of the participants' demographics and experience data were almost homogenous for both groups (lab vs NE). Just over half of the participants indicated that they were native English language speakers (52.0%). The non-native speakers rated their English level as either 'fairly fluent' (16.7%) or 'moderate fluency' (31.3%).

Half of the participants* were undergraduates (50%), 41.7% were master's students, and 8.3% were studying for PhDs. These percentages were the same for the two groups. Sixty-

* The main study was carried out by students who were currently studying at UEA as either undergraduates or master's or PhD students. UEA graduates were excluded as they no longer had a UEA email address and, based on the experimental criteria mentioned in Section 5.2.2.5, they were not accepted for participation.

six of the participants (45.8%) indicated that they used digital libraries 'occasionally' or 'monthly' in their normal practice before participating in this experiment. Thirty participants (20.8%) reported that they used digital libraries 'frequently' or 'fortnightly', 12 (8.3%) used digital library websites 'always' or 'weekly or semi-daily', and 36 (25%) reported rare usage of digital library websites.

No technical issues were reported by the participants for any task in either experimental condition. None of the participants had previous experience with any of the digital libraries' websites. Ninety participants had previous experience using Amazon (93.75%). Fifty-eight participants indicated that they had 'occasionally' used Amazon (40.3%), 54 (37.5%) 'always' used it, 22 (15.3%) rarely used it, and 6 (6.6%) had 'never' used it. The distribution of experience with Amazon.co.uk for the entire sample was similar to the distribution in the subsample of each testing environment setting. The independent t-test confirmed that the experimental groups did not differ in their self-rated experience with amazon, $p = 0.436$.

5.2.3.3 Analysis Approach

The analysis approach for this study was based on three sequential phases. The first phase involved screening the data and usability testing data to match the data (Figure 5.9). This matching allowed us to explore the data based on the screening data as in the previous section and investigate the influences and/or relationships between the user characteristics used in the screening data on the usability testing data. The usability testing data were composed of usability testing outcomes and other testing data (Table 5.7). After the data were checked and statistical controls applied if needed, the processes depicted in Figure 5.10 were carried out.

Table 5.7. Components of Usability Testing Data

<table>
<tr><th colspan="6">Usability Testing Data</th></tr>
<tr><th colspan="5">Usability Testing Outcomes</th><th rowspan="2">Other Testing Data</th></tr>
<tr><th colspan="2">Perceived usability</th><th colspan="3">Performance</th></tr>
<tr><td>SUS Scores</td><td>Usability Issues</td><td>No of Successful Completions</td><td>Page Views</td><td>Time on Tasks</td><td>Time Elapsed on Questions</td></tr>
<tr><td colspan="4"></td><td colspan="2">Time Elapsed on the Entire Test</td></tr>
</table>

The first statistical analysis carried out was to investigate whether the task complexity influenced the usability testing outcomes in the different experimental conditions (lab vs NE). The time taken to complete all four tasks was measured.

Repeated measures MANOVA analysis confirmed that there was no interaction effect between task type and experimental condition (lab vs NE) ($V = 0.059$, $F(12, 83) = 0.434$, $p = 0.945$, $d = 15$ (very large), $1- \beta = 1$ (perfect), (Figure 5.11). The results for the other usability testing outcomes showed that the testing outcomes did not differ between the different experimental conditions (lab vs NE) for each task with a certain difficulty level (Table 5.8). Thus, the focus was on the between-subjects variation (the two different experiments; Figure 5.12).

Participants' characteristics were found to have no effect on any of the usability outcomes for each experimental group. However, a multivariate significant difference was found between English language levels and the elapsed time for the entire test ($\lambda = 0.917$, $F(12,46) = 3.247$, $p = 0.02$, $d = 0.01$, $(1- \beta = 0.6)$.

This difference is not induced by the experimental conditions (lab vs NE), as no significant interaction was found between the experimental conditions and the English language level ($\lambda = 0.158$, $F(12, 46) = 0.328$, $p = 0.980$, $d = 5$, $1- \beta = 1$). Table 5.7 shows the time taken to complete the entire test, which is composed of Time on Tasks and Time on Questions. Thus, to verify the influence of English language level on the time taken to complete the entire test, we investigated whether an influence was incurred by Time on Tasks by applying a univariate independent one-way ANOVA.

The result showed that the difference between English language level was found for the Time on Questions, $F(2,27) = 16.00$, $p < 0.00$. $d = 1$, $1- \beta =1$, but not for Time on Tasks, $p = 0.655$, $d = 0.2$, $1 - \beta = 0.8$.

To determine the influence of English language level on the time taken to complete the questions, we applied Tukey-way post-hoc analyses and found that participants who considered themselves to have a moderate English language level took significantly longer to answer the questions ($p < 0.000$) than those who rated themselves as 'fairly fluent' and 'fluent'.

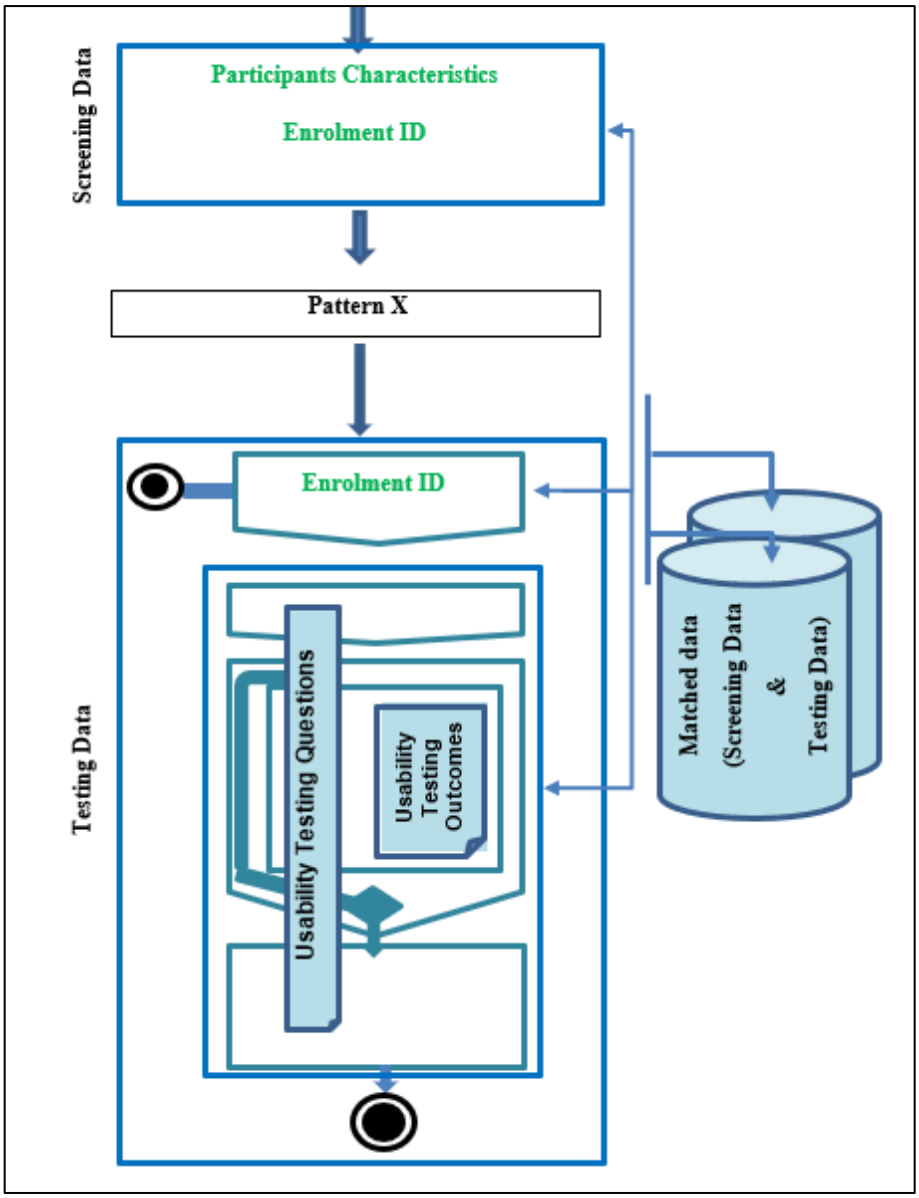


Figure 5.10. Study analysis approach, matching data.

Table 5.8. Interaction Effect of Task Complexity on Usability Testing Outcomes with Regards to the Experimental Conditions (lab vs NE)

	Perseus		**UDL**		**arXiv**		**Amazon**		
Experimental conditions	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	
Descriptive / **Time on Task**	**Mean, (SD), N**								**Interaction Effect**
Time on Task	81.17, (16.19), 48	83.40, (15.12), 48	115.96, (32.54), 48	114.62, (32.17), 48	283.58, (84.52), 48	315.27, (102.81), 48	97.02, (21.39), 48	100.38, (23.626), 48	F (1.327, 123.799) = 2.304, p = 0.123
Page Views	3.41, (0.500), 48	3.56, (0.558), 48	3.76, (1.046), 48	3.69, (0.856), 48	5.85, (1.329), 48	5.36, (1.199), 48	3.62, (0.652), 48	3.58, (0.649), 48	F (2.478, 161.692) = 1.659, p = 0.188
SUS Scores	78.021, (8.613), 48	78.698, (8.902), 48	78.906, (11.048), 48	79.792, (10.364), 48	45.990, (17.387), 48	45.625, (17.240), 48	81.615, (9.488), 48	81.927, (9.358), 48	F (2.272, 213.564) = 0.050, p = 0.965
Usability Issues	0.67, (0.753), 48	0.69, (0.776), 48	0.71, (0.713), 48	0.69, (0.689), 48	1.00, (0.583), 48	1.04, (0.544), 48	0.19, (0.394), 48	0.19, (0.394), 48	F (2.744, 257.904) = 0.050, p = 0.980

The p-value for significance is 0.05

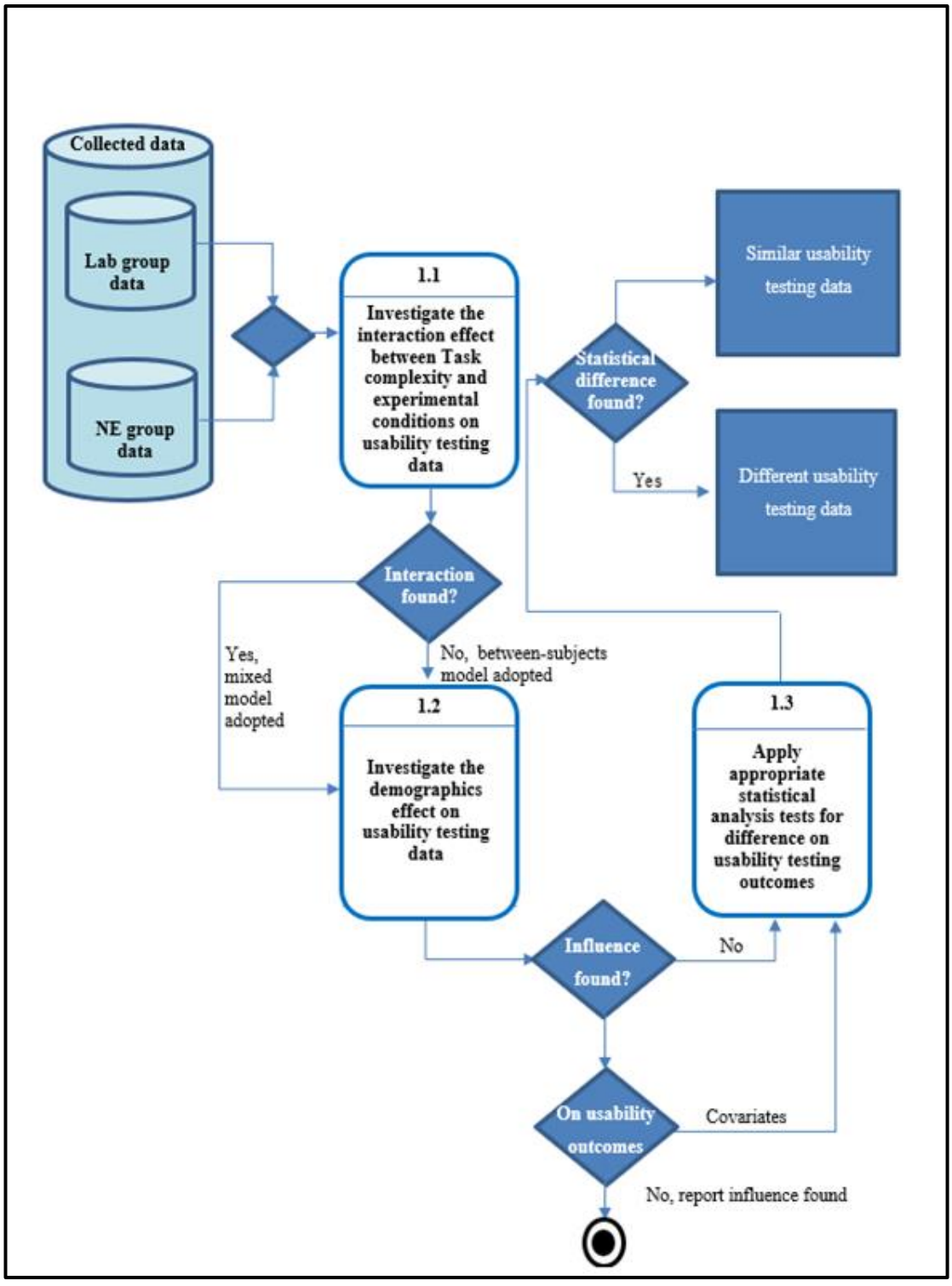


Figure 5.11. Study analysis approach, statistical control activities' flow diagram.

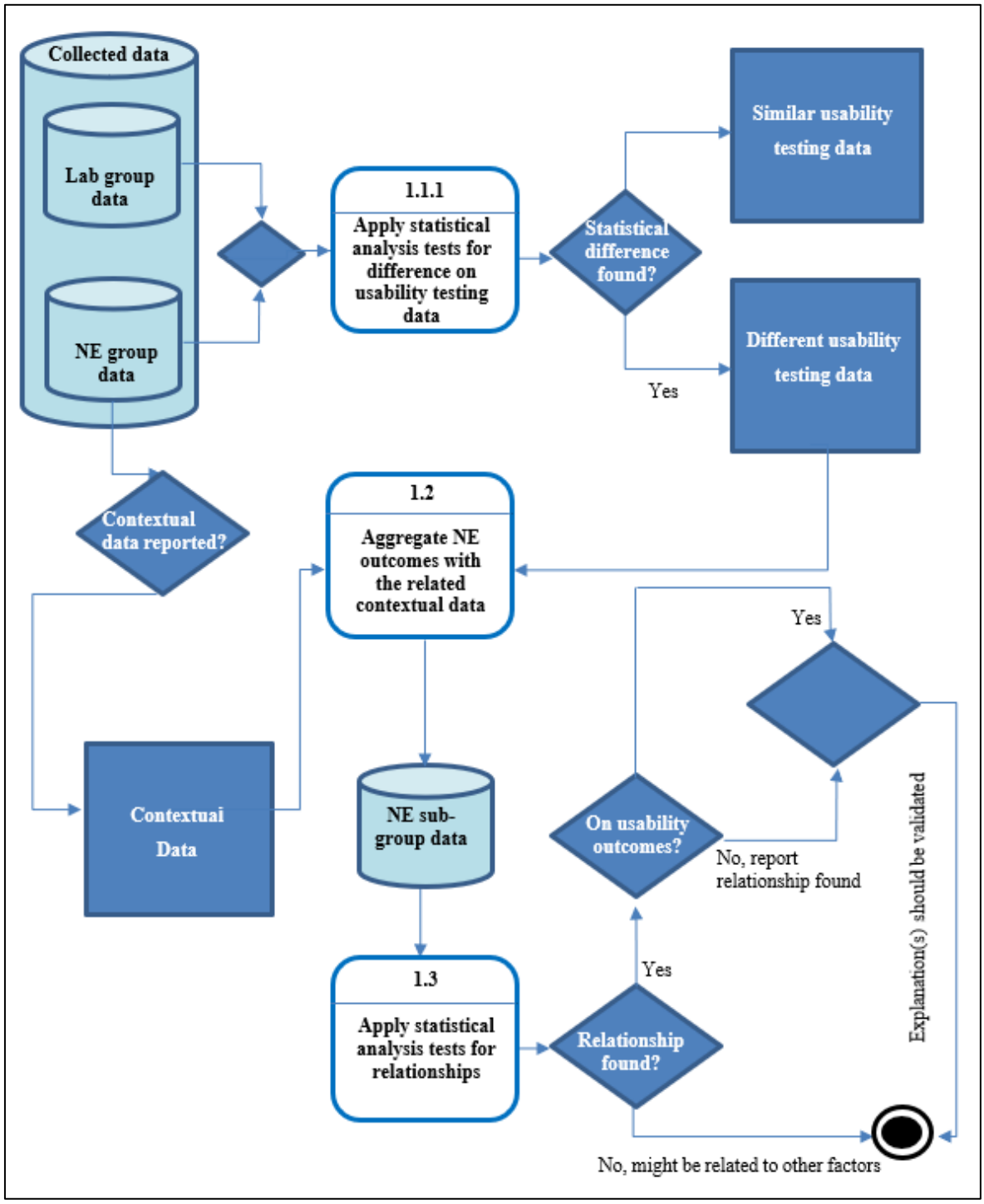


Figure 5.12. Study analysis approach, formal statistical analysis activities flow diagram.

5.2.3.4 Usability Testing Outcomes

Usability testing outcomes are represented as performance outcomes and perceived usability reports. Performance outcomes are represented by Time on Tasks, Page Views and Number of Successful Completions, while perceived usability reports are represented by SUS scores and participants reports about usability issues (Table 5.9). The statistical analyses showed

that no differences existed in the usability testing outcomes between the two experimental conditions in terms of performance and perceived usability outcomes (see Table 5.9 for a summary of the findings).

- **Performance**

With respect to performance outcomes, a multi-factorial ANOVA indicated a non-significant between-subjects difference between the experimental conditions (lab vs NE) for Time on Tasks, F (1,94) = 2.296, p = 0.133; d = 2.1 (large), 1-β = 1 (perfect).

For the Page Views, the multi-factorial ANOVA test showed a non-significant difference between the groups, as follows: F (1, 68) = 0.977, p = 0.327, d = 0.119 (small effect).

For the perceived usability, a mixed 4 × 2 multi-factorial ANOVA test was applied, which showed a non-significant difference between the groups, i.e. F (1, 94) = 0.094, p =0 .670, d = 0.03.

With respect to usability issues, another mixed 4 × 2 multi-factorial ANOVA test was applied, showing a non-significant difference between the groups, i.e. F (1, 94) = 0.094, p = 0.670, d = 0.03.

To investigate whether experimental conditions (lab vs NE) were associated with the successful task completions rate, Fisher's exact test was applied. The results indicated that no significant association was observed between the testing environment and the successful rate task completion rate for Perseus: (p = 1.000), φ = 0.000 (Phi coefficient of no effect).

Similarly, no association was found between the testing environment and the successful rate task completion rate for the UDL task based on Yates' continuity correction analysis: Yates' (1) = 0.000, p = 1.000, φ = 0.030 (very minor effect). Similar results were obtained for the arXiv task, Yates' (1) = 1.555, p = 0.212, φ = 0.148 (minor effect). No statistical test could be conducted for Amazon because all tasks were successfully completed for both experimental conditions (Figure 5.13).

5.2.3.5 The Control Task Outcomes

We revisited the usability testing outcomes with Amazon across the two experimental conditions (lab vs NE). Table 5.10 shows that no significant differences were found between the two experimental conditions for all usability testing outcomes.

This means that if we control for task complexity (or if we use only one task in the usability evaluation), a significant difference is unlikely between the usability testing outcomes for the two environmental conditions.

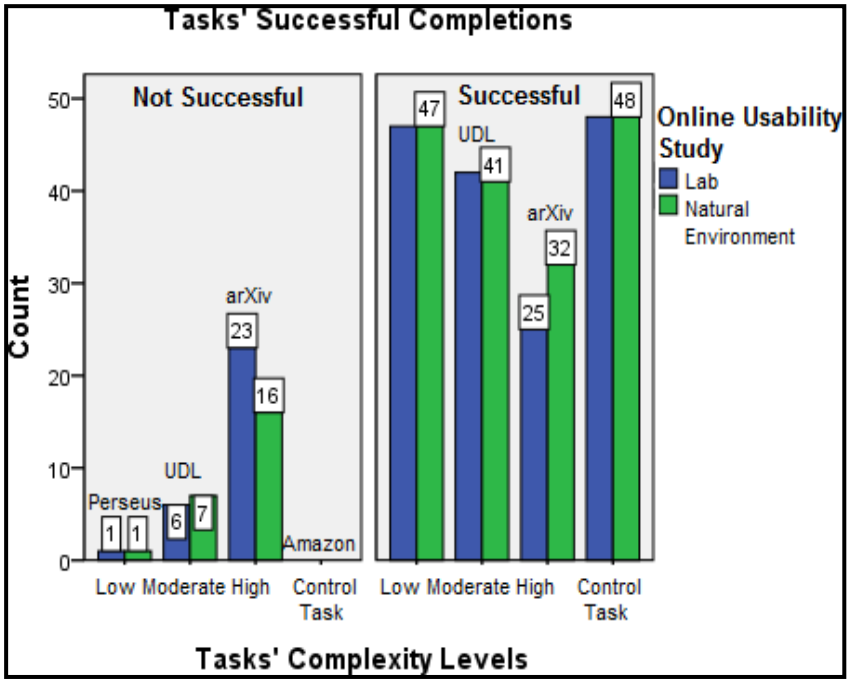


Figure 5.13. Tasks completions for each experimental condition.

5.2.3.6 Type of Contextual Factors

Contextual factors were only by reported by NE participants as no external distractions were allowed and the systems were controlled in the lab environment.

- **Distractions**

Thus, the data presentation covers only the sub-group of NE participants that reported distraction events (interruptions and multitasking).

- Interruptions

  Only 10 participants of the NE group (20.8%) indicated that they experienced interruptions during the test. However, no more than two interruptions were experienced by one participant during the test session. Seven (14.6%) of the participants who experienced interruptions only experienced one interruption during the entire test session, and three participants (6.3%) experienced two interruptions during the entire test session.

Table 5.9. Usability Outcomes for Each Task, and All Tasks for Both Experimental Conditions

	1st Task		2nd Task		3rd Task		4th Task		All tasks		Statistical Test for All Tasks
Experimental conditions	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	**Lab**	**NE**	
Descriptive	**(Median: SD)**										
Time on Task	(89.7: 30.7)	(120.50: 53.4)	(96.40: 47.53)	(107.73: 48.18)	(222.90: 123.87)	(23.10: 15.1)	(90.67: 21.24)	(147: 59.1)	(507.6: 140.21)	(620.1: 245.5)	F (1,94) = 2.296, p = 0.133; d = 2.1 (large), 1 - β = 1
Page Views	(3.5: 0.93)	(3.83: 1.724)	(4.90: 1.38)	(4.17: 0.51)	(7.70: 3.974)	(5.4: 3.55)	(5.20: 1.989)	(4.50: 1.77)	(23.10: 5.859)	(19.15: 6.08)	F (1, 68) = 0.977, p = 0.327; d = 0.119
Successful Completions	8	10	9	17	2	1	9	16	23	19.5	NA
Perceived Usability	(2.20: 1.23)	(1.95: 0.99)	(3: 1.32)	(3.10: 1.071)	(4.50: 0.966)	(4: 1.16)	(1.20: 000)	(1.11: 0.32)	(2.73, 0.74)	(2.49, 0.57)	F (1, 94) = 0.094, p = 0.670, d = 0.03.
Usability Issues	(0.8: 1.14)	(0.6: 0.8)	(0.3: 0.9)	(0.40: 0.7)	(1.6: 1.08)	(1.7: 1.4)	(0.3: 0.5)	(0.30: 0.5)	(3: 2.2)	(3.5: 1.8)	F (1, 94) = 0.094, p = 0.670, d = 0.03.

Table 5.10. Control Task's Usability Outcomes among the Experimental Conditions

	Experimental Conditions			
	Lab	**NE**	**Statistical Test**	***p*-value**
Descriptive	**(mean: SD)**			
Time on Task	97.02, 21.393	100.38, 23.626	t-test	$p = 0.468$,
Page Views	3.58, 3.56	0.647, 0.616	t-test	$p = 0.872$,
Successful Task Completions	NA	NA	NA	NA
SUS scores	81.61, 9.488	81.93, 9.358	t-test	$p = 0.871$,
Usability Issues	0.19, 0.394	0.19, 0.394	Mann–Whitney	$p = 1.000$,

The *p*-value for significance is .05.

Most of the participants (6, 60%) who indicated that they experienced an interruption during the test performance indicated that this was a direct in-person conversation. This type of interruption accounted for six (50%) of the reported interruptions, and receiving calls accounted for 16.6%. One instance was reported of hearing other people's conversation nearby, receiving text messages via text applications, receiving broadcast via chat applications and other social activities, e.g. 'watching over kids' (1, 8.3%), respectively.

- Multitasking

Slightly more than half (25, 52.1%) of the participants in the NE group reported that they had other applications or tasks open on the computer they were using to perform the test (e.g. an office application).

The number of tasks (other than the test's tasks) open during the test session was not more than three, and only one participant reported that they had four applications/programs open when performing the test.

Of the 52.1% of the participants who had applications or programs open, 15 (31.3%) had only one program open, 7 (14.6%) had two programs open, 2 (4.2%) had three programs open and only 1 participant (2.1%) indicated that they had four programs open while performing the test.

All the participants who admitted they had other applications or tasks open had their email open. Email comprised 25 (62.5%) of all reported multitasking events. Based on the adopted Cohen (1980) classification between interruption and multitasking, as detailed in Section 2.3.3, email notifications were considered a multitasking event as they would pop up on the screen if they were set up that way by the participant; hence, we reasoned that we would consider it a multitasking event. Seven (17.5%) of the reported multitasking events were having another website open, three (7.5%) were with Facebook, three (7.5%) were with Skype, and two (5%) were with office applications (word processors and spreadsheets).

However, most of the participants (21, 43.8%) reported that they did not look at these programs or applications, and thus they could not be considered a distraction influence. Figure 5.14 shows the distribution of distraction events reported by participants in the NE group (a) for interruptions and (b) for having other programs open.

- Apparatus

Most of the participants (40, 83.3%) in the NE group reported that they had used devices with large screens (e.g. laptops or PCs). Only four participants (8.3%) reported that they had used devices with medium screens (e.g. medium handheld devices, such as iPads and tablets), and only four participants (8.3%) reported that they had used devices with small screens (e.g. small handheld androids and smartphones; Figure 5.15(a)).

With regards to the internet connection speed, most (42, 87.5%) of the participants in the NE group reported that they had used a relatively fast internet connection speed (e.g. the UEA network or a fast connection somewhere else). Five participants (10.4%) indicated usage of a relatively medium internet connection speed (e.g. a modem), and only one participant (2.1%) indicated that they had used a relatively low internet connection speed (e.g. mobile or dial-up; Figure 5.15(b)).

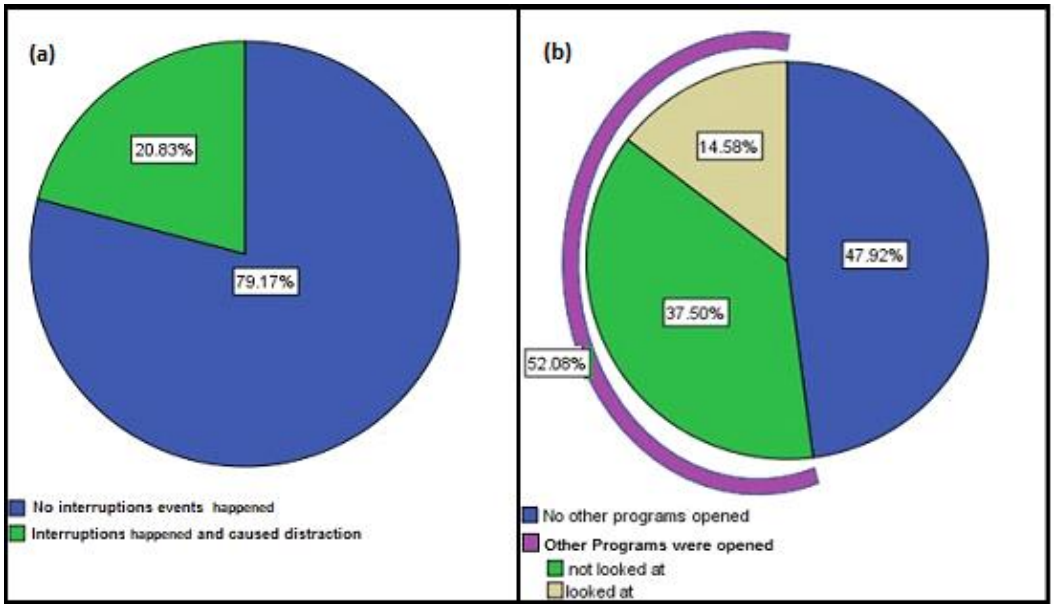


Figure 5.14. Frequency of distraction events reported during experimental usability testing in the NE for (a) interruptions and (b) other programs open.

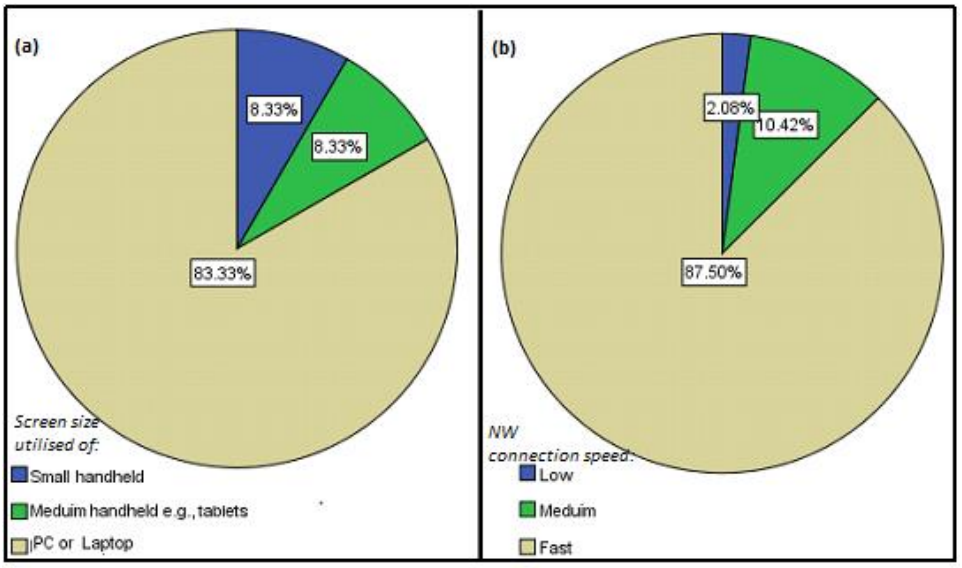


Figure 5.15. Frequency of system types used and internet connection speed in the NE group.

5.2.3.7 Relationship between Usability Testing Data and Contextual Factors

As shown previously, no differences were found in any of the usability testing outcomes between the experimental conditions (lab vs NE). Based on the analysis approach adopted for this study (Figure 5.10), if a difference is found in the usability testing data, we need to first aggregate the NE testing data outcome where the difference is found with the related

contextual data and apply statistical analysis tests to investigate the differences and/or relationships. Therefore, we first needed to explore which components of the usability testing data were different among the two experimental conditions (lab vs NE). As indicated in Section 5.2.2, usability testing data is composed of usability testing outcomes and other testing data, which is represented in the time taken to answer the questions. In other words, this time refers to any time elapsed during the entire test except for the time recorded for each task. We will call it Time on Questions from now on. Most of the previous literature in RAUT which acknowledged differences in the time measurement referred to the time measurement as Time on Tasks; however, when reviewing those studies, we realised that the time reported is mostly the time taken for the entire test, including the testing tasks. Nevertheless, we also check the Time on Tasks to enable a comparison.

Following the analysis approach, we first investigate whether Time on Questions and Time on Tasks differ between the two experimental conditions. To do this, we applied a MANOVA model using Wilks' lambda test to simultaneously examine the influence on the Time on Questions and Time on Tasks while accounting for English language level. The results indicated a significant effect of the interaction between the experimental conditions and participants' English language levels on the time scores: $\lambda = 0.887$, $F(4, 178) = 2.754$, $p = .030$, $d = 0.248$ (medium) and 1-β err prob = 0.44. Table 5.11 shows the mean and SD of Time on All Tasks and Time on Questions with respect to the two experimental conditions.

Table 5.11. Statistics for Time on All Tasks and Time on Questions in The Online Usability Study (lab vs NE)

	(Mean: SD)	**Experimental conditions**
Time on tasks	(577.73: 109.102)	***Lab***
	(613.67: 122.882)	***NE***
Time on Questions	(859.90: 249.540)	***Lab***
	(1175.62: 425.346)	***NE***

However, a subsequent post-hoc test showed that this significant difference affected *only* Time on Questions and not Time on All Tasks. The t-tests showed a non-significant effect on Time on All Tasks, $F(1,90) = 1.52$, $p = .221$, but a significant effect on Time on Questions, $F(1,90) = 31.71$, $p < .001$, $d = 0.91$ (large) and 1-β err prob = 0.99 (very strong).

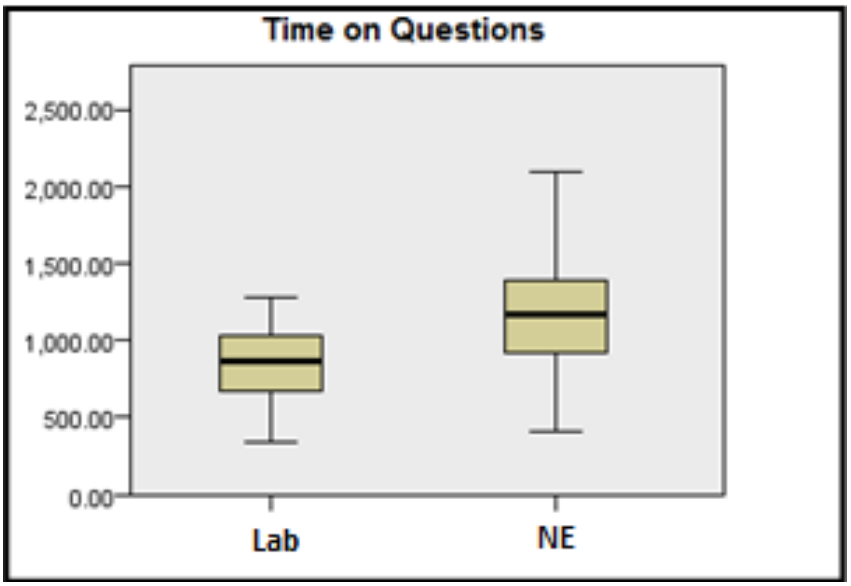


Figure 5.16. Mean values of Time on Questions (in seconds) for both experimental conditions.

No significant results were found for the effect of the testing environment combined with the English language level on Time on All Tasks. However, a significant result was found for the same effect on Time on Questions, F (2,90) = 4.414, p = .015 (Figure 5.16). As the distribution of the participants with the different English Language Levels were homogeneous for the two experimental conditions, it is conceivable to say that Time on Questions was influenced by the participants' English language level regardless of the experimental condition.

Now we realise which component of the testing data ensured the difference between the two experimental conditions (lab vs NE), we select the NE participants' data where they have reported distractions and contextual factors and apply statistical analysis tests to investigate the differences and/or relationships.

Regarding interruptions, 20.8% of the NE participants indicated that they experienced interruption(s) while performing the test. Because of the extremely unbalanced results for the groups (20.8% and 79.2%), the exact test was used.

The *p*-values generated using the Monte Carlo technique* of the Mann-Whitney test showed that a significant difference existed in the time scores between the participants who indicated they were distracted by interruptions and those who were not on Time on Questions: U = 92.00, $p = 0.012$, Z = -2.488. The Monte Carlo technique guarantees with 99% confidence that the true *p*-values were contained within the (0.009-0.014) range.

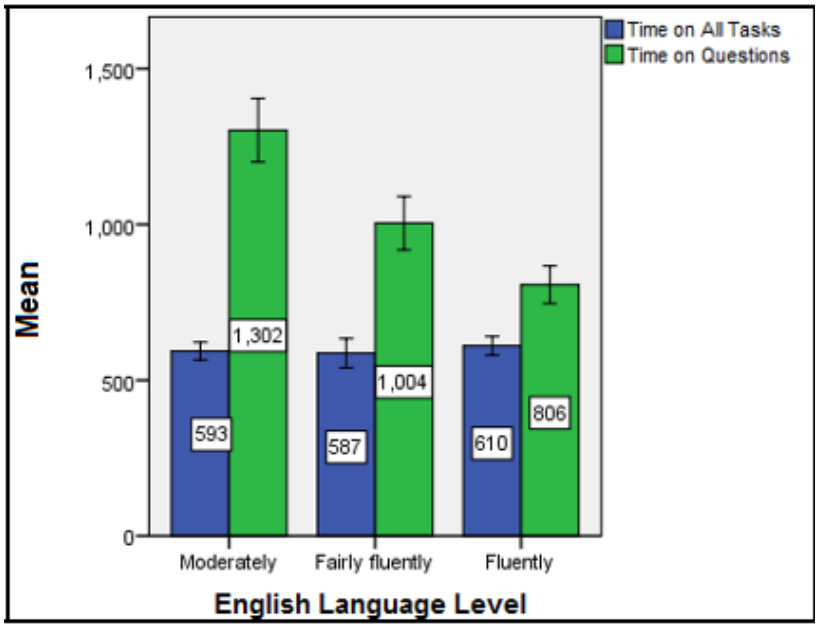


Figure 5.17. Mean values of time on total tasks and time on questions (in seconds) with respect to English language level.

With respect to screen size, using the Monte Carlo technique, the *p*-values generated using the Monte Carlo technique of Kruskal-Wallis test showed that there is a significant difference, with regard to device screen size, on Time on Questions only, $\chi 2$ (2, n = 48) = 17.946, $p = 0.000$. The Monte Carlo technique assures with 99% confidence that the true *p*-value is contained within (000-000) range. However, as we have three groups associated with either the 'small', 'medium', or 'large' device screen size, we still do not know which groups are significantly different from one another. Thus, a follow-up Mann-Whitney U test was applied with a Bonferroni adjustment to the alpha values with each group-pair comparison to control for Type 1 errors (Tabachnick and Fidell, 2013, p. 52). This adjustment involves dividing the alpha by the number of comparisons to be made. As we

* Monte Carlo technique was used instead of the exact test as the sample size of the NE group, 48 participants, is not properly suited to the exact test.

have three pair comparisons, the alpha value was 0.017. The results showed that there was a significant difference in Time on Questions between participants who were using small and large devices screens (U= 0.00, p = 0.00, Z = -3.27) an those who were using medium and large devices screens (U = 0.00, p = 0.000, Z=-3.26).

With respect to Internet connection speed, excluding participants with a slow Internet connection*, the *p*-values generated using the Monte Carlo technique of Mann-Whitney test showed that there is a significant difference on Time on Questions, *only*, between who were utilising medium Internet connection from those who were utilising fast medium Internet connection. The Monte Carlo technique assures with 99% confidence that the true p-value is contained within (000-000) range. Refer to Table 5.12 which show a summary of the results of the tests applied to the Time on All Tasks and Time on Questions with respect to the contextual factors.

Table 5.12. Median and Number of Participants Who Reported Interruptions During Task Performance and Those Who Did Not, With Respect to Time Scores

		Usability Testing Data Component	Median	Number	Statistical Test	True *p*-value Range
Interrupted?	***Yes***	**Time on All Tasks**	611.00	10	Mann-Whitney U tests	**(0.688-0.712)**
	No		658.50	38		
	Yes	**Time on Questions**	1097.50	10		**(0.009-0.014)***
	No		1630.00	38		
Multitasking	***Yes***	**Time on All Tasks**	658.00	7	Mann-Whitney U tests	**(0.419-0.444)**
	No		607.00	41		
	Yes	**Time on Questions**	1251.00	7		**(0.253-0.276)**
	No		1107.00	41		
Screen Size	***Small***	**Time on All Tasks**	599.50	4	Kruskal-Wallis test U tests	**(0.359-0.383)**
	Medium		672.00	4		
	Large		611.00	40		
	Small	**Time on Questions**	1729.50	4		**(0.000-0.000)***
	Medium		1840.50	4		
	Large		1041.00	40		
Connection Speed	***Medium***	**Time on All Tasks**	636.00	5	Mann-Whitney U tests	**(0.872-0.889)**
	High		617.00	42		
	Medium	**Time on Questions**	1859.00	5		**(0.000-0.000)***
	High		1065.00	42		

The previous tests showed that Time on Questions was significantly influenced by contextual factors. To answer RQ4 more concisely, we subsequently ran a correlational analysis to determine whether the variance on Time on Questions related to the contextual

* The exclusion was decided as only one participant indicated the usage of slow connection.

factors. A significant correlation was found between English Level and Time on Questions: $r_s = -0.693$, $p < .001$, between Interruptions and Time on Questions: $r_s = -0.343$, $p = 0.008$, and between Connection speed and Time on Questions: $r_s = -0.552$, $p < 0.001$ (Table 5.13).

We performed a multilinear regression to examine how much of the variance in Time on Questions for the NE participants was explained by contextual factors. A significant regression model, using the Stepwise* method ($F(3, 44)) = 22.628$, $p < 0.001$) predicted 61.2% of the sample outcome variance (Adj. R^2 0.580). Three predictors – lower English language level ($\beta = -247.922$, $t = -5.127$, $p < .001$), higher interruption occurrence ($\beta = 48.272$, $t = 2.373$, $p = 0.022$) and lower connection speed ($\beta = -223.169$, $t = -2.119$, $p = 0.040$) – were significantly associated with longer question times. Two other predictor variables (having other tasks running and display size) were excluded from the model (Table 5.14).

Table 5.13. Spearman's Correlation Significant Results for Contextual Factors with Time on Questions

	Contextual Factors		**Time on Questions**
Spearman's rho	English Level	**Correlation Coefficient**	-0.693**
		Sig. (1-tailed)	0.000
		N	48
	Interruptions	**Correlation Coefficient**	0.343**
		Sig. (1-tailed)	0.008
		N	48
	Connection Speed	**Correlation Coefficient**	-0.552**
		Sig. (1-tailed)	0.000
		N	48

**. Correlation is significant at the 0.01 level (1-tailed).

* The variation in the dependant variable examined in series of steps in a form of a nested models, where the researcher has a rationale for having multiple steps of regression and for choosing which variable is the first variable. Most restricted model would be the one in the first step and the most general one is the one in the last step (Mayers, 2013).

Table 5.14. Multiple Linear Regression (Stepwise) Analysis for Time on Questions

Predictor Variable	**R^2**	**Adj. R^2**	**R^2/change**	**F**	***p***	**Gradient**	**t**	***p***
Model	0.607	0.580		22.628	<.001			
English Level			0.506			-247.922	-5.127	<0.001
Interruptions			0.061			48.272	2.373	0.022
Connection Speed			0.040			-223.169	-2.119	0.040

5.3 Discussion

In this comparative explanatory study, we investigated the differences in the usability testing outcomes in terms of participants' performance and subjective reports. We examined the contextual factors experienced and reported by participants in the NE group and identified whether a relationship exists between the usability testing outcomes and the contextual factors reported in terms of participants' performance and subjective reports.

The study met its first objective of taking into account the issues found in the previous exploratory study and applying the suggested design features (Figures 5.4 and 5.8). The second objective was also achieved as the design of this comparative study was enhanced and several design and statistical controls were applied, as discussed in Sections 5.2.2.7 and 5.2.3.3. The third objective to investigate the contextual factors reported by remote participants during their usability testing session was also achieved (see Section 5.2.4.3). The fourth objective to investigate the difference in usability testing outcomes was also met, as participants' performance and subjective ratings were statistically compared between different testing environment settings and related findings were reported (Sections 5.2.41 and 5.2.4.2). The fifth objective was also met by investigating the relationship between the contextual factors reported by participants and the differences in the usability testing outcomes provided in Section 5.2.4.4.

Having achieved the study's objectives, we discuss the findings with relation to RQ2:

> RQ2: Does usability testing data performance during usability testing in the (remote) natural environment differ from that of participants in a lab environment?

The findings showed that no differences existed with regard to usability testing outcomes between the NE and lab environments. However, a significant difference was found for Time

on Questions. Usability testing outcomes varied on the task level, whereas Time on Questions comprised the total time elapsed for the tasks, excluding the time consumed on the tasks. This finding replicates our exploratory finding and agrees with Greifeneder (2011), who stated that 'people in the natural environment needed statistically more time to complete the test' (p. 312). Given those findings, would the rigorous design of this study and the sampling technique used emphasise that Time on Questions is an indicator of contextual factors? Consider the scenario in which a NE participant have experienced distractions and wanted to report them, would they have taken longer to answer the question(s) about whether they had been distracted? A conflicting scenario might take place when there was a longer Time on Question(s) because that participant was reporting usability issues. That is, Time on Questions could be used as an indicator of an unusual interaction or experience during the usability testing. Whether it related to contextual factors should be further investigated by determining the reason for their existence and determining whether a relationship or correlation exists. RQ3 and RQ4 aim to fill this gap:

> RQ3: What contextual factors do remote participants experience during their usability testing session?
>
> RQ4: How do the contextual factors influence the users' outcomes during usability testing?

We based our classification of distractions as interruptions and multitasking on the definition and classification of Cohen (1980; Section 2.3.3). Many participants reported having other tasks running (multitasking); however, they indicated that they did not look at them while performing the task(s). Interruptions were less frequent but had a greater influence based on the participants' feedback. That is, with multitasking, participants decide whether to switch between tasks or carry out tasks, while interruptions are intrusive and beyond the decision-maker's control. This explanation might interpret participants' negative feedback regarding interruptions despite a lower frequency than multitasking during usability testing. This explanation also agrees with Cohen (1980) about interruptions and multitasking and indicates that participants prefer to perform the tasks and choose not to multitask even if other applications are open in the background. Participants might consider that usability testing is a finite specified task which will be carried out in one session and, hence, they might prefer to avoid being distracted during their performance. However, this explanation differs slightly from the findings and explanations reported in workflow studies. Again, the

nature of task in usability testing might explain the difference. Hence, it is important to be aware of distractions in the context of usability testing, as participants cannot control their occurrences.

With respect to connection speed, we operationalised the options to low, medium, and high NW connection speed. Device screen size was operationalised into small, medium, and large, depending on the type of computing/communication machine used to access the test. Data showed that participants of the NE group chose to access the test using larger sized computing devices (e.g. PCs, laptops, notebooks and tablets) and a more reliable network connection technology (UEA network or WIFI technology), and they used a 3G mobile connection technology when using a mobile phone. These findings indicate that participants prefer to optimise their experience when taking part in the usability testing and choose computing devices with bigger display screens and faster network connection technology if they can. However, these inferences remain unconfirmed, given the absence of participants' feedback to confirm our inferences.

However, a correlational analysis offers a better understanding and appreciation of what happened during the NE testing sessions. The correlational analysis showed a significant correlation between English level and Time on Questions, interruptions and connection speed. The regression analysis showed that the variance on Time on Question is explained, mainly, by English language level, followed by frequency of interruptions and connection speed.

Chapter 6: Interrupted Tasks Influence on Usability Testing

6.1 Overview

The previous chapter presented the empirical explanatory study which aimed to answer the second, third and fourth RQs. The previous explanatory study's findings indicated no differences in the usability outcomes between the lab and NE groups. However, a significant difference was found in Time on Questions between the two environments. Further analyses showed that English language influenced Time on Questions in both testing environments. With respect to the NE group, Time on Questions was found to be influenced mainly by whether the performance was interrupted and the connection speed. The previous study gave valuable explanations of usability testing outcomes and data in the NE group.

However, in practice, usability practitioners should care only about Time on Tasks, since this metric reflects the time a user requires to perform a given task. Time on Questions is not meant to reflect users' real experiences with a product, since it deals primarily with the time taken to answer self-reported questions. In other words, it is not a usability testing outcome. From a different perspective, we still cannot be sure that interruptions *cause* the negative effect on usability testing outcomes, as acknowledged in most RAUT literature, and we did not detect whether this influence exists. We reasoned in the discussion of the previous study that it is likely that participants are more likely to interrupt their performance during question time rather than task time. That is, we argue that usability practitioners are more concerned with the data yielded by users out of the usability testing rather than the time needed to report on the testing experience. Hence, these issues should be considered and addressed in a further study designed for such purpose. This is therefore the main objective of this validation study.

The remainder of the chapter is organised as follows: Section 6.2 describes the objective and presents the general design of this experimental study and discusses the OUUT tool used for the data collection. Section 6.3 presents the discussion.

6.2 The Experimental Validation Study

6.2.1 Study Objectives

This validation study investigates the cost of the interrupted tasks in usability testing with respect to usability testing performance. This study answers RQ5:

> RQ5: What is the cost of interrupted users' performance in usability testing to usability practice?

To answer RQ5 using the OUUT, this study seeks to meet the following objectives:

- Validate the previous study's findings in terms of the relationships found between interruptions and time measurements.
- Design an experiment which controls all confounding variables to isolate the factor to be investigated: interruption influence.
- Investigate the differences in usability testing performance between the interrupted tasks and the non-interrupted task performance.
- Investigate the differences between the task-load incurred by the interrupted tasks and the non-interrupted task performance.
- Investigate the interruption cost in terms of how the task(s) performance would be influenced by interruptions.
- Obtain insights about which type of interruption is the most disruptive for participants to perform the task.

By designing and conducting this experimental study, we aim to meet the above objectives and answer the RQ.

6.2.2 Study Design

To answer RQ5, we design an online usability study that applies RAUT, which is accessible by participants in a controlled lab environmental setting, where all the confounding factors are controlled, except for the interruptions.

The effect is presented as a cost, which refers to the time taken to reorient towards task performance. Existing literature suggests that interruptions result in longer completion times (e.g., Czerwinski et al., 2000; Bowman et al., 2010; Kirschner & Karpinksi, 2010). Furthermore, while the English language and NW connection speed could be controlled in a

practical online usability study, interruptions cannot. Consequently, to investigate the effects of interruptions on participants' performance, we controlled experimentally for English language, NW connection speed and display size.
The participants in the previous study reported external interruptions in the form of phone calls, instant messaging and in-person conversations. To isolate the variables of interest, interruptions in a lab environment were operationalised and simulated during the testing session.

Passive observations were carried out using a passive recording tool, as no physical observations were made to back up the performance data. Therefore, recordings of video, audio or the participant's screen were obtained. In addition, the entire session was streamed in real time to enable the test-facilitator (the researcher) to apply the interruptions systematically.

Our primary variable of interest was the total time taken to perform the test tasks. The total time needed to complete each block of tasks was automatically recorded by the OUUT. The frequency of interruptions was applied systematically. The time spent on the interruption was manually recorded by the test facilitator, who observed the tasks' performance without being present in the same room. The time to perform the tasks was computed as total time to perform task minus time spent on interruptions. If the time to perform the task was higher with an interruption, then this could indicate that extra time was needed to perform the task after an interruption.

Additionally, errors, defined as the number of deviations from the perfect path to accomplish a certain task, represented testing outcomes that were translated into the actual performance. Errors are different from participants' feedback regarding usability issues in the previous study, which the participants reported in their own words. We argue that an interruption is more likely to influence the efficiency of how users accomplish the tasks, and consequently, they might be more vulnerable to committing errors. Errors were recorded manually by calculating the number of deviations from the perfect task performance path using the screen recordings of participants' task(s) performance.

Subjective reports were measured by a modified NASA Task Load Index (TLX). We used the NASA TLX as it can be adjusted to have five rating scales: time pressure, effort, mental

demand, stress and frustration. Participants were required to rate these factors on the standard NASA 20-point scale in a way that did not interfere with task performance or influence time measurement. That is, they were required to use the NASA TLX paper and pen forms.

To obtain insights into which type of interruption was the most disruptive for participants during task performance, the task participants' subjective feedback was collected. This was attainable as the participants performed the experiment in a lab and were interviewed after completing the experimental tasks. The participants were asked about the extent to which they were for some reason disturbed, which prevented them from fully immersing themselves in the experimental task. They were also asked which interruption type was the most disruptive and why? Participants' feedback was manually recorded.

6.2.2.1 OUUT: Loop11

Loop11, discussed previously in Section 4.2.2.1, was used to administer the experimental tasks for the participants online. Loop11 was used because it can automatically record the time per each task and record the screen to review participants' performance and identify their errors. The questions facility in Loop11 was used to instruct participants to move between the tasks' blocks and the NASA TLX paper and pen forms.

The collected data were transferred directly into a spreadsheet file. URLs of the pages visited for each task in each test session were stored as textual entries in the spreadsheet file. Data were automatically collected, updated and transferred into the spreadsheet file. Logged performance in terms of visited URLs and clickstreams were automatically recorded and saved using Loop11. These records were then utilised for analysis.

6.2.2.2 Experimental Design and Tasks

A repeated measures experimental design was used. The within-subjects independent variable was the interruption sources with four levels: No interruption source (B: baseline), Phone interruption (Ph), Instant Messaging interruption (IM) and Physical interruption by person (Pr). These simulated interruptions simulated the sources of external interruptions reported by participants in the previous study. The order of interruption sources was fully counter-balanced.

One test object was used for this experiment – the Durham University Library Website (www.dur.ac.uk/library). The home page of this website includes a search engine positioned in the middle of the page and a number of links for various options that are standard for most academic library websites, such as conducting searches, booking a study room and booking a library computer. The website has a mixed base interface that combines navigation and reading. All information on the site is available only in English. The library website of Durham University was chosen as the test object for this study because it did not require participants to sign in as students to perform searching tasks.

The sample consisted of students from UEA as they are considered typical target users for such a website. The searching tasks were similar to those used in the previous study, as one of the main objectives of this experimental study is to validate the findings from the previous study. In addition, the flow in performance where an interruption takes place is more relevant if a problem-solving task is carried out (e.g. searching tasks). We argue that participants might be eager to solve the task and reorient it after an interruption occurs if it is a problem-solving task rather than another task type (e.g. structured task). For each interruption source, participants had to perform four tasks; each group of four tasks is referred to as a 'task block'.

Task blocks are designed to be similar but not identical. Identical tasks per task block were avoided because, even if counterbalancing was applied during the experimental setting, the participants might not perform the tasks honestly to find the desired information and it would be easy for them to perform the tasks; hence, the interruption might not have a considerable effect on their performance. We thus opted for problem-solving tasks with different attributes as the experimental testing tasks.

The tasks were similar to where they should be positioned in every task block, such that Task A.1, Task B.1, Task C.1, and Task D.1 were similar, Task A.2, Task B.2, Task C.2 and Task D.2. were similar and so on for the third and fourth task blocks, which ultimately made Task A block, Task B block, Task C block, and Task block D similar. However, the tasks within each block were different so that Task A.1 differed from Task A.2, Task A.3 and Task A.4. Differences among the block's tasks were incurred by designing the tasks to be accomplished using different performance paths, such as key information, search feature, limit to function and information, for each task within the block (Table 6.1).

Table 6.1: Task Block Design for The Validation Study

Task Block		Task ID: 1	2	3	4
Similar	**A**	A.1 Author	A.2 Shelf mark + Limit to function	A.3 Title + subject	A.4 Title, Author, Material type language
	B	B.1 Title	B.2 Subject + Limit to function	B.3 Author + Note	A.4 Subject, Material type, Years range, Language
	C	C.1 Subject	C.2 Author + Limit to function	C.3 Note + title	C.4 Note, Note, Section
	D	D.1 Shelf mark	D.2 title + Limit to function	D.4 Author + subject	D.4 Title, Subject, Years range

Different

In addition, the task blocks were similarly mentally demanding and time-consuming, for example, calculating how many clicks or pages were required to achieve the required information or solving the tasks and determining how difficult they were to perform.

We developed several tasks designs and made several design reviews, which involved asking some participants to carry out the designed tasks every time to check the time per task and determine how mentally demanding they were. The last design review showed that the task blocks were equally demanding and required a similar time to complete. For this last design review, we ran two mini pilot tests.

Time on Task was measured in seconds and automatically recorded by Loop11, in which the designed tasks were administered. Mental Load was measured using the Subjective Mental Effort Questionnaire (SMEQ), which is made up of one scale with nine labels ranging from 'Not at all hard to do' to 'Tremendously hard to do' (see Figure 6.1). After the participant finished each task, they were given a pen and paper showing the items of SMEQ as millimetres above the baseline, and the scale ranging from 0 to 150 (Figure 6.1).

Using the scale, the participants were asked to draw a line through a vertical scale to indicate the amount of effort they needed to invest to execute the task. SMEQ is reliable and easy to use (Zijlstra, 1993; Kirakowski and Cierlik, 1998) and it correlates highly with task completion time, completion rates and errors (Sauro and Lewis, 2012, p. 214). In addition, SMEQ shows good sensitivity for small sample size compared to other post-task questionnaire measurement scales (e.g. SEQ, UME; Sauro and Dumas, 2009).

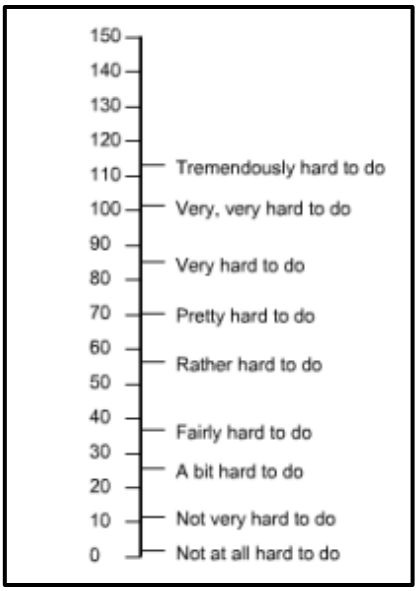


Figure 6.1: SMEQ (Source: Sauro and Dumas, 2009).

- **Mini-pilot 1**

This pilot sought to check whether the time required to perform each task and the mental load required to execute each task within the task block was similar among participants. To meet this purpose, we used a mixed within-between subjects' statistical design, in which each participant carried out only one task block, such that for that task block, they were required to carry out four individual tasks which form that task block. Thus, the between-subject variables are Task Block and Task ID, and the dependent variables are Time and Mental Load. The Task Block varied on four levels, A, B, C and D, and the Task ID varied on four other levels: Task 1, Task 2, Task 3 and Task 4. Six participants were recruited, and each participant was given a £5 Amazon.co.uk voucher after completing their tasks. Table 6.2 shows the Time and Mental Load scores towards time, which varied by Task Block and Task ID. For the Time scores, a mixed 4 × 4 multi-factorial ANOVA indicated a non-significant between-groups difference for the time required to perform the task blocks, F (3, 20) = 0.074, p = 0.974. Time on Tasks within each task block was found to be significantly different, F (1.476, 29.517) = 11.885, p = 0.001*. Regarding whether the time spent on corresponding Task ID within blocks was similar, we found no interaction of task blocks

*All Task ID pairs are significantly different, except for T1 vs T4 and T2 vs T4.

with the time spent on individual tasks, F (4.428, 29.517) = 0.219, p = .938. We examined whether the total time per whole block was similar. An independent one-way ANOVA indicated that no significant difference was evident for time spent on the four different task blocks, F (3, 20) = 0.74, p =0 .974.

For the Mental Load scores, a mixed 4 × 4 multi-factorial ANOVA indicated a non-significant between-groups difference in the Mental Load ratings scores given to the task blocks, F (3, 20) = 0. 289, p = .833. Additionally, the Mental Load rating scores given for task within task blocks were found to be significantly different, F (1.956, 39.12) = 1456.52, $p < 0.001$. We examined whether the Mental Load score given to each corresponding Task ID within the blocks was similar. We found no interaction of task blocks with the Mental Load score given to the individual task, F (5.86, 39.12) = 0.551, p = 0.802. We checked the total Mental Load required per whole block. An independent one-way ANOVA indicated that time spent on tasks block was not significantly different among the four different task blocks, F (3, 20) = 0.289, p = 0.833.

- **Mini-pilot 2**

The previous mini-pilot examine the consistency in the time taken to complete the task blocks and in the incurred mental load. In this mini-pilot, we investigated whether the same participants performed the four tasks blocks consistently. We applied a within-subjects' statistical design, which required each participant to carry out the four test blocks. The task blocks were counterbalanced using the ordered Latin squares technique. Thus, every individual task within a certain task block was compared with the corresponding task within the other task blocks. For example, the time score of a participant per task for A.1, B.1, C.1 and D.1 should be compared with A.2, A.3 and A.4 in the other tasks' blocks. This process was also applied to obtain the Mental Load scores.

The task blocks were administered online using Loop11, which automatically recorded the time taken to complete each task. After completing each task, participants were instructed to use the Loop11 interface to answer the SMEQ questions. Then, participants were instructed to go back to the Loop11 interface to perform the next task. Eight participants were invited to carry out this pilot, receiving a £7 Amazon.co.uk voucher upon completion. A Kruskal-Wallis test found no significant differences in time for individual tasks for the task blocks: H (3) = 2.108, p = 0.550 (Table 6.3).

Table 6.2: Mini-pilot 1: Task Block Design: Time and Mental Load Scores for Each Task within Task Blocks by Participant

	Time				**Mental Load**			
	Task Blocks				**Task Blocks**			
	A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
	Mean, (SD), N	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**
Task 1	81.50, (22.17), 6	97.83, (54.16), 6	76.16, (19.47), 6	76.16, (22.13), 6	11.16, (3.18), 6	10.83, (1.47), 6	11.00, (2.60), 6	10.33, (1.03), 6
Task 2	144.66, (69.75), 6	142.16, (71.51), 6	130.83, (73.16), 6	126.83, (77.39), 6	105.00, (10.48), 6	111.83, (14.06), 6	106.66, (10.80), 6	105.00, (10.48), 6
Task 3	157.50, (82.23), 6	153.33, (80.93), 6	171.33, (65.36), 6	145.33, (86.37), 6	115.00, (10.48), 6	114.83, (13.18), 6	117.16, (13.87), 6	115.50, (13.47), 6
Task 4	103.66, (33.39), 6	112.83, (33.65), 6	103.16, (34.50), 6	111.33, (34.87), 6	135.00, (10.48), 6	134.83, (9.80), 6	140.66, (10.93), 6	135.00, (10.48), 6
All Tasks	487.33, (175.90), 6	506.16, (162.38), 6	481.50, (168.98), 6	459.66, (184.10), 6	85.58, (6.28), 6	87.50, (5.69), 6	87.95, (6.18), 6	86.58, (5.82), 6

Table 6.3: Mini-pilot 2: Task Block Design: Time and Mental Load Scores for The Task Blocks Carried Out by a Participant

	Time					**Mental Load**				
	Task Blocks				**Kruskal-Wallis Test**	**Task Blocks**				**Kruskal-Wallis Test**
	A	**B**	**C**	**D**		**A**	**B**	**C**	**D**	
	Mean, (SD), N	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**		**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	**Mean, (SD), N**	
Task 1	86.13, (22.13), 8	95.63, (48.01), 8	80.00, (20.22), 8	73.63, (19.35), 8	H (3) = 2.108, p = 0.550.	10.63, (2.87), 8	10.88, (3.27), 8	10.63, (3.66), 8	10.88, (2.80), 8	H (3) = 0.905, p = 0.824.
Task 2	166.13, (77.91), 8	167.13, (73.49), 8	162.00, (78.34), 8	159.25, (80.41), 8	H (3) = 2.108, p = 0.550.	104.38, (14.50), 8	105.00, (14.14), 8	106.63, (11.18), 8	107.25, (12.37), 8	H (3) = 0.346, p = 0.951
Task 3	154.63, (91.70), 8	177.75, (83.75), 8	179.38, (71.27), 8	171.75, (89.52), 8	H (3) = 2.684, p = 0.443	136.75, (9.57), 8	137.00, (11.38), 8	137.25, (10.44), 8	136.75, (10.40), 8	H (3) = 0.284, p = 0.963
Task 4	100.13, (28.97), 8	106.25, (31.35), 8	104.88, (34.79), 8	104.13, (32.54), 8	H (3) = 0.158, p = 0.984	89.13, (5.19), 8	89.13, (5.33), 8	89.00, (5.34), 8	89.13, (5.34), 8	H (3) = 0.333, p = 0.954
All Tasks	507.12, (139.65), 8	546.75, (83.80), 8	526.25, (66.13), 8	508.75, (132.12), 8	H (3) = 1.200, p = 0.753	340.87, (15.58), 8	342.00, (16.86), 8	343.50, (16.20), 8	344.00, (17.00), 8	H (3) = 1.423, p = 0.700

6.2.2.3 Experimental Conditions

We simulated a real usability testing session to determine the outcomes based on the previous explanatory study. Consequently, the experimental conditions for this study are representations of the various interruption sources reported in the explanatory study (Table 6.4).

Table 6.4: Experimental Conditions

Experimental Conditions	**Interruption Source**
B	Baseline condition with no source for interruptions
Ph	Phone
IM	Instant Messaging
Pr	Person (conversation with a physically present person)

The frequency of interruptions might impact task performance (Lee and Duffy, 2015). Hence, the interruptions frequency was fixed for each interruption source. That is, the interruptions frequency was set to two minutes after the start of each task block based on the pilots and as suggested by extant literature (Gillie and Broadbent, 1989; Mark et al., 2008). During the experiment, the experimenter adjusted the length of the interruptions to make the interruption durations as equal as possible across all interruption context conditions, which is ≈ 2 minutes based on the pilots.

Three questions were designed using mental arithmetic problems as cognitive process tasks. Our choice was justified by Lee and Duffy (2015, p.138), who stated that 'cognitive process task requires more mental demands to complete than a motor skill task, it is likely that the former is more susceptible to interruptions than the latter'. See Table 6.5 for the transcript of the questions. The questions were designed to be similar in complexity yet different in the approach required to work out the answer.

Another design review was carried out with 18 volunteers who rated their experience during the interruption while performing a single task block (A), where each was exposed to a certain interruption question applied through certain interruption sources (Ph, IM, or Pr) on the questionnaire shown in Figure 6.1. The Cronbach's α for internal reliability was 0.873, indicating satisfactory consistency among the three interruption questions.

In the actual experiment, we did not associate each interruption source (except for baseline) to certain questions; rather, we counterbalanced the questions with interruptions. The intention was to control for the possibility of mixing the interruption source types, whether they were delivered by phone, IM or in person, with the mental demands incurred by the cognitive process for the question.

6.2.2.4 Study Advertisements

Multiple methods were used to recruit participants for the experiment.

- Official email using UEA mailing lists
  An official email was designed and circulated to the students of multiple schools at UEA. The email included the study's purpose, importance, guarantee of data confidentiality, consent information, test duration, incentive amount and method of receiving the incentive.
- Flyers were disseminated in UEA union building seating areas and cafés.
- A4 posters were placed on the bulletin boards containing identical content to the flyers.
- Social media: A Facebook page was used, containing identical content to the flyers and posters.

6.2.2.5 Experimental Controls

The inclusion criteria were as follows:

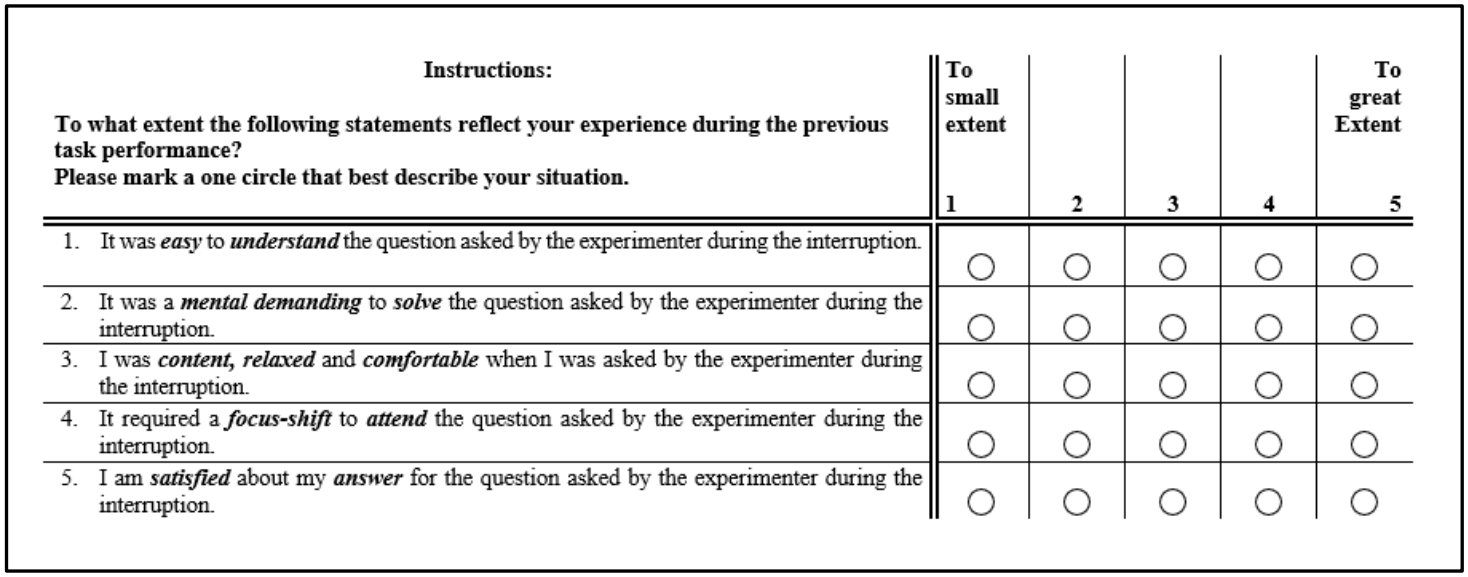
- UEA current students with a valid UEA email address
- Never participated in any usability testing before
- Have used smartphones to receive calls
- Have used smartphones to use instant messaging applications (e.g. WhatsApp)

6.2.2.6 Ethical Clearance

The data collection materials, including Zoho (www.zoho.com/), Loop11, Camtasia, Skype, Participant File, TLX and the interview guide, were ethically approved before starting the experimental procedures. Before seeking ethical approval, several pilot tests and redesigns were then carried out. Once the experiment was fully designed, all the documentation, including the required participant reassurances, screenshots of the study design materials and informed consents were submitted to the Ethical Approval Committee of the Computing Science School at UEA. A few adjustments were required to obtain final approval for the designs of the data collection and advertisements (Appendix A.CH6).

Table 6.5: Transcript for the Questions Used for Interruptions

Question	Question transcript	Design specification
Q1:	Could you tell me how you travelled to the experiment session location today? Could you work out how many weeks left until the summer term, which starts on the 16th of July?	Opening question: Arithmetic (Work out a time point in the future)
Q2:	Could you tell me how you found out about this experiment? Could you work out how many weeks have you been in UEA since the start of the spring semester, which started on the 15th of January?	Opening question: Arithmetic (Work out a time duration in the past)
Q3:	Could you tell me how you contacted me to show your interest in participating in this experiment? Suppose that you have been offered a summer employment between the 22nd of July and the 9th of September, how many weeks will you have been at work?	Opening question: Arithmetic (Calculate time duration based on the difference of two time points)



Instructions: **To what extent the following statements reflect your experience during the previous task performance?** **Please mark a one circle that best describe your situation.**	**To small extent** **1**	**2**	**3**	**4**	**To great Extent** **5**
1. It was ***easy*** to ***understand*** the question asked by the experimenter during the interruption.	○	○	○	○	○
2. It was a ***mental demanding*** to ***solve*** the question asked by the experimenter during the interruption.	○	○	○	○	○
3. I was ***content, relaxed*** and ***comfortable*** when I was asked by the experimenter during the interruption.	○	○	○	○	○
4. It required a ***focus-shift*** to ***attend*** the question asked by the experimenter during the interruption.	○	○	○	○	○
5. I am ***satisfied*** about my ***answer*** for the question asked by the experimenter during the interruption.	○	○	○	○	○

Figure 6.2: Design review: a questionnaire to rate the level of interruption caused by the designed questions.

6.2.2.7 Experimental Protocol

In this experiment, we simulated the interruptions that were likely to take place in users' NE, and we collected the required measurements and designed the experimental tasks, procedures, and statistical design and controls.

The number of participants needed for the experiment was a multiple of four because the experimental design is a within-subjects design where four exposures (conditions) were applied. As counterbalancing was applied, we had to have all possible permutations needed to collect the required data. Consequently, we aimed to recruit $16 \leq X \leq 48$, where X is the number of participants.

After advertising the study, students expressed their interest in participating in the study via the email address provided in the study's advertisements. Then, the online experimental controls were applied. The participants received a screening questionnaire to complete, which was designed using Zoho (Appendix A.CH6). After screening the participants, the selected participants received an email confirming their acceptance and including a link to the study schedule on Doodle (https://doodle.com), where the scheduling process was carried out. Participants were granted access to Doodle using their UEA email provided in their emails. As the participants were already registered in the study schedule on Doodle, they were able to assign themselves an hour occupancy between 8:00 AM and 5:00 PM during a three-week period. Only one selection was allowed per IP access. A confirmation email was sent to the participants, including information about the location and time of the test.

The experiment was conducted in a quiet lab at UEA's Computing Science School. The lab was divided into two rooms. The participant performed the test in the bigger room, and the researcher observed from the small room, which had a door with a glass window. However, the glass window was very small and was only used as a back-up for the streamed data obtained through Skype.

When the participants arrived, they were welcomed to the test room, where they were given the test instruction document, which informed participants about what could and could not be done during the test. For example, participants could not use their personal mobile phone during the experimental session and they could not open any other window but the Loop11 window. They were also informed that the experimenter *might* contact them during the

session for any reason, and if so, they should attend to these contacts as soon as they happen. They were asked to use only the smartphone provided on the participant desk to answer phone calls or WhatsApp messages from the experimenter. The smartphone provided contained only the experimenter's contact in the phone book and WhatsApp app. The smartphone was connected to the UEA network to enable online messaging through WhatsApp. The instruction document included explanations about Loop11 interface and functionality and gave explanations about the searching tasks. The explanation of searching tasks guaranteed the minimum level of awareness of how to conduct searching tasks using online dynamic websites. Once the participant finished reading the instructions, they were asked to sign the informed consent form, which was in a pen and paper format.

Meanwhile, the test moderator (the researcher) opened the corresponding study based on the task blocks' order and according to the counterbalancing scheme. The test moderator opened the Camtasia tool in the background to record the screen. Participants gave their consent for recording the screen or video, but they were unaware if it was happening to avoid any possible influence. The test moderator then assigned a Participant File and a session ID. The participant was handed the participant file, which included the informed consent that they should sign to start the experiment. The participant was then directed to use Loop11, which guided them through the session. The participant was asked to start performing the experimental tasks using the laptop provided on the desk when they felt ready. Task blocks were administered online using Loop11, which automatically recorded the time taken to complete each task. The sequence of interruptions to be applied on that session were already predetermined considering their types, (Ph, IM, or Pr), and the corresponding questions. The same case was applied to the pattern of the questions to be asked. The time consumed per interruption and the resumption time were recorded manually by the test moderator in the Session Log File, which was assigned the same session ID.

Another laptop was used with a Skype application running, such that a Skype video call enabled the camera and the microphone to stream the participant's performance and activities during the experimental session to the experimenter's machine. The video streaming of the sessions enabled the experimenter to observe the participant's reaction to the interruption when they returned to the task performance after the interruption and when they started the new task blocks. The call opened by the experimenter to enable the streaming

process was only an audio call with the microphone off, so the participant was not influenced by this setting (see Table 6.6 for the systems used).

Table 6.6: Devices and Apparatus Used in The Validation Study

Device used	Purpose	Hardware	Software
Computer	To enable participants to perform the experimental tasks.	UEA Laptop Type: Toshiba	Loop11 using Google Chrome
Bowser	To enable participants to perform the experimental tasks.	Utilised in UEA Laptop Type: Toshiba	Google Chrome
Built in Cam 1	To video stream the experimental tasks performance and test in real time.	Mac Air (A) Built in Cam	Skype Video caller on Mac Air (A) device
Built in Cam 2	To enable the experimenter to receive and monitor the video streaming of the experimental tasks performance and test in real time.	Mac Air (B) Built in Cam	Skype Video caller on Mac Air (B) device
Smartphone 1	To enable the experimenter to perform the phone and instant messaging interruptions.	iPhone 7, Phone (A)	• iPhone Caller • WhatsApp
Smartphone 2	To enable the participant to receive and respond to the phone and instant messaging interruptions.	iPhone 6, Phone (B)	• iPhone Caller • WhatsApp

After every two minutes of the start of a new task block, the test moderator applied the corresponding interruption (asking a certain question in a certain interruption form) and started manually recording the time consumed during the interruption. Task resumption was considered once the participant clicked or moved the mouse or pressed a key of the keyboard.

After completing the performance of each task block, participants were instructed using Loop11 interface to carry out the NASA Task Load Index (TLX) on a paper format that was included within the Participant File. Then, participants were instructed to go back to Loop11 interface to perform the next task blocks (see Appendix A.CH4 for more details).

Once the participants finished their experimental session, they were interviewed to clarify some issues about their performance and their experience during the experimental session. The answers for the interview questions were documented in the Session Log File. Finally, they were thanked and given their £10 token incentive.

6.2.3 Study Analysis

6.2.3.1 Data Preparation

As data were collected through Loop11, the spreadsheets for the different study versions (based on tasks order patterns) were retrieved and associated with the interruption log data.

The time and date were automatically recorded for each data entry in the spreadsheet file, which enabled the association with the Participant File and Session Log File to be done. Then, the manually recorded data in both files were populated to their related automatically recorded data in the spreadsheet file generated by Loop11. The Time on Tasks score was updated, excluding the time for interruptions from the corresponding Time on Block for that session.

Then, the SPSS 25.0.0.0 data statistics tool was used to read the data and perform the required statistical analyses. Using the SPSS tool, the data were coded properly. Quality checks were carried out on the data.

After completing the data preparation, the data were checked to see whether they had a normal distribution. If the data were found not to be normally distributed, data transformation techniques were used, if applicable, to transform the data, as detailed in Chapter 3. Then, appropriate statistical analysis tests were selected based on the data nature and the type of the RQ to be answered.

Participants' feedback obtained in the interview was transcribed verbatim into word processing files for analysis. During the transcription process, the transcriptions were checked for accuracy and the data formatted and organised to facilitate the analysis.

6.2.3.2 Data Exploring

Forty-eight participants participated in the study, 26 females and 22 males. Of those, 75% were native English UEA university students, 4.16% were bilingual, and 20.83% were non-English speakers who scored more than 6.5 on the IELTS test. The majority (62.5%) of the participants were aged 35-44 years, followed by 29.2% who were aged 25-34 years and 8.3% (4 participants) who were aged 35-44. Of the participants, 62.5% were undergraduates, 25% were doing a master's, and 12.5% were doing PhDs. Most of the participants (37.5%) majored in applied sciences, 35.4% in social sciences, 20.8% in art and humanities, and 6.3% in medicine and health sciences. All the participants had been using the Internet for more than five years; 66.7% had been using IM for at least five years; 18.8% had been using IM for at least three years but less than five years, 14.6% had been using IM for more than one year but less than or equal to three years. Participants were given a £10 token for their participation.

6.2.3.3 Analysis Approach

The analysis approach for this study was based on four sequential phases. First, data matching was performed between the data collected by Loop11 and data retrieved from the Participant File and Session Log File. The data were combined and matched appropriately in one single tabular form to be readable by SPSS as a data source file.

Second, the quantitative analysis using SPSS was carried out using the related statistical tests to answer the study question, and the results of the tests were described.

Third, the qualitative analysis was applied to the secondary source of data – the interview data. There is no systematic procedure that all qualitative researchers follow (Creswell and Plano Clark, 2017). Thus, the researcher should identify the best approach to address the RQs (Creswell and Plano Clark, 2017). In this study, we followed the method which helped to organize our data. Thus, all the participants' feedback from the interview was read to develop a general understanding of the data, and memos or themes were coded to record broader categories of information, such as codes or themes. A qualitative codebook was then developed.

Fourth, the quantitative and qualitative strands were integrated such that a mixed methods analysis was applied, as the design of this study added the qualitative data collection (the questionnaire) into the experiment to include the personal experiences of the participants. This enabled us to demonstrate how qualitative data augmented the experiment's results, for example, by using a joint display that can present the integration of the experimental and qualitative results.

6.2.4 Study Findings

6.2.4.1. The Cost of Interrupted Task in Usability Testing

- **Quantitative analysis results**
    - Performance

      With respect to Time on Tasks, the repeated measures ANOVA indicated that a significant difference in Time to Perform Task according to the forms of the interruptions applied, F (2.31, 108.59) = 5.210, $p < 0.05$. A post-hoc Bonferroni

analysis indicated that participants took significantly longer to perform the task in the Pr condition than in the B condition ($p < 0.05$). No significant difference was observed between the B and Ph conditions, between the B and IM conditions, and between the B and Ph condition vs the IM and Pr conditions. These findings were represented by a medium effect, d = 0.3, 1-β = 0.99, with very high power.

A significant difference was found in the number of errors participants made across interruption forms, as indicated by the repeated measure ANOVA, F (2.43, 114.55) = 18.220, $p < 0.001$. A post-hoc Bonferroni analysis indicated that participants made significantly more errors in the IM condition than in B condition ($p < 0.001$) and Ph conditions ($p = 0.001$). In addition, significantly more errors were committed in the Pr condition than in the B ($p < 0.001$) and Ph conditions ($p < 0.05$). However, no significant differences were found between B versus Ph conditions and between the IM and Pr conditions. These findings were represented by a large effect, d = 0.6, 1-β = 1.00 with very high power. Table 6.7 shows the descriptive data and Table 6.8 summaries the statistical results.

Table 6.7: Descriptive Data of Performance Measurements

	Time on Tasks	**Errors**
Interruption forms	**Mean, (SD)**	**Mean, (SD)**
Baseline/No Interruption (B)	19.27, (14.47)	15.8, (13.81)
Phone Interruption (Ph)	25.72, (13.72)	29.47, (17.66)
Instant Messaging Interruption (IM)	37.81, (14.97)	41.77, (13.58)
In-Person Interruption (Pr)	35.62, (17.70)	41.35, (16.78)

Table 6.8: Differences for Time on Tasks and Number of Errors across Interruptions along Significant Post-hoc Bonferroni Pair-wise Comparisons

	Statistical Test	*p*-value	Effect Size	Statistical Power	Significant Results of Post-hoc Bonferroni Analyses
Time on Tasks	F (2.31, 108.59) = 5.210	$p < .05$	d = 0.3 medium effect	1-β = 0.99 very high power	Pr vs B ($p < .05$)
Errors	F (2.43, 114.55) = 18.220	$p < .001$	d = 0.6, large	1- β = 1.00, perfect	IM vs B ($p < .001$)
					IM vs Ph, ($p < .001$)
					Pr vs B, ($p < .001$)
					Pr vs Ph, ($p < .001$)

- Workload

A repeated measures analysis showed that mental workload was rated as significantly different across interruption forms, F (2.589, 121.6) = 101, $p < 0.001$. A post-hoc Bonferroni analysis showed that mental load was rated as significantly different between the B condition versus the IM condition ($p < 0.001$), and versus the Pr condition ($p < 0.001$). In addition, mental load was rated as significantly different between the Ph and IM conditions ($p < 0.001$) and versus the Pr condition ($p = 0.021$). However, no significant difference was observed between mental load ratings between the B and Ph conditions or between the IM and Pr conditions. This was represented by a large effect, d = 0.6, 1- β = 1.00, with very high power. Tables 6.9 and 6.10 show the descriptive data of the other task load measurements and the statistical findings, respectively.

Table 6.9: Descriptive Data of Workload Measurements

Interruption forms	**Workload Measurements**				
	Mental load	**Time Pressure**	**Performance**	**Effort**	**Frustration**
	Mean (SD)	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**
Baseline/No Interruption (B)	19.27 (14.47)	11.14 (11.77)	16.97 (16.26)	19.68 (17.30)	14.37 (12.61)
Phone Interruption (Ph)	25.72 (13.72)	21.66 (16.92)	11.59 (9.53)	27.08 (15.43)	27.70 (19.26)
Instant Messaging Interruption (IM)	37.81 (14.97)	34.16 (14.45)	6.562 (8.603)	40.20 (14.08)	40.83 (16.76)
By Person Interruption (Pr)	35.62 (17.70)	26.56 (17.92)	6.146 (9.47)	35.00 (18.62)	39.58 (17.82)

- **Qualitative Analysis Results**

  - Participants' feedback

All the participants reported feeling uncomfortable during the interruptions. One participant stated, 'Everything was difficult; usually I handle that, but the last task when nobody called me, it was a bit better'. Another said, 'I felt so nervous when I was asked those questions', and another participant said, 'Yes, those questions were really disturbing'.

The participants described the Pr (29, 60.41%) and IM (19, 39.58%) conditions as the most disruptive interruptions, and some of those participants (11, 22.91%) claimed

that they found the Ph interruption the least disruptive. Regarding the Pr condition as the most disruptive interruption, several participants indicated that they were unable to get back directly to their previous state of mind afterwards. For example, one participant said, 'I had more difficulties concentrating on the task after you popped in and started asking me questions'. Another participant stated, 'It took me a bit of time to remember what I was specifically doing after you left'. One said, 'That task when you asked me here took forever for me to find what it asked about'. Another participant indicated, 'I hardly remembered what I had to do and how to complete that task', and another explained, 'I stopped for a little bit before I continued working on the task'.

Some of the participants who indicated that they were highly distracted by IM interruptions more than the other types also stated some interesting points. One participant stated, 'I am not quite sure about my performance for that task when you texted me, though!'. Another participant said, 'I found it more disruptive to handle the text messages, as every time I thought I would not get a new message and I was about to resume the task a new message came'. One said, 'It was hard to shift my mind between the task and the messages'.

Some participants indicated that they tried to focus on both tasks (the primary and interrupting task) but they could not; for example, one participant said, 'I found when you messaged me on WhatsApp, it was really annoying because I was trying to focus on the tasks and answer you at the same time!' Another participant said, 'I was trying to get the task right, and I was so focused on the task, but at the same time the messages got my mind away, really!' Some participants indicated that they were under stress: 'It is quite stressful to answer the messages and try to resume the task'.

Table 6.10: Qualitative Analysis Results

Coded responses	**No of occurrence**
Faulty performance	4
Require higher mental load	11
Stress	4
pressure	6
Poor performance	2
Hard effort	3
Frustrated	5
Time to resumption	13

Another one added ‘To answer your messages… that was stressful’. Some other participants acknowledged being under pressure trying to handle the messages, ‘But I couldn’t do that. The messages were so frequently sent’, and another one stated, ‘I was under a pressure to reply to the messages as I wanted to return to the task as quickly as I could’. Another participant acknowledged making much effort to perform the task, ‘I tried hard to shuffle between the messages and the task’. Another participant indicated their frustration: ‘I felt frustrated trying to answer your messages’.

- **Mixed-methods analysis findings**

Tables 6.12 and 6.13 show the relationship between the experiment outcomes and participants’ experiences, illustrating the combination of numeric values and textual qualitative data in a single display.

The verbatim responses of the codes shown in Table 6.11 were placed beside the corresponding measurement (usability testing outcomes) used in the experiment along with the corresponding qualitative result. This way, we can see that the participants’ comments augment the quantitative results by giving more explanation through their descriptions of their experience towards the outcome.

Table 6.12 shows the participants’ responses indicating that IM is the most disruptive interruption, while Table 6.13 shows those that suggest that Pr is the most disruptive interruption. The findings that 29 of the participants (60.41%) found Pr the most disruptive interruption, while 19 (39.58%) found IM the most disruptive interruption support this study’s quantitative findings (see Tables 6.8 and 6.10).

6.3 Discussion

In this validation study, we investigated the cost of the interrupted task performance on usability testing performance. The study met its first objective because it validated the previous study’s findings in terms of the relationship found between interruptions and time measurements (Chapter 5). The second objective was also achieved, as this study controlled all confounding variables, detailed in Section 6.2.2, enabling us to explore the influence of the interruptions. The third objective was to investigate the differences in usability testing

performance between the interrupted task and non-interrupted task performances, which was achieved (see Section 6.2.4.1). The fourth objective to investigate the differences between the task-load incurred by the interrupted tasks and the non-interrupted task performance and related findings was also met (see Section 6.2.4.1). The fifth objective was met by investigating the interruption cost in terms of how the task(s) performance was influenced by the interruptions (see Section 6.2.4.1). The sixth objective to obtain insights about which type of interruption was the most disruptive for participants to perform the task was also achieved (Section 6.2.4.1).

As the study achieved its objectives, we discuss the findings with relation to RQ5:

RQ5: What is the effect 'the cost' of interrupted users' performance in usability testing to usability practice?

The findings showed that a significant difference existed in the performance outcomes between interrupted and non-interrupted task performance, depending on the forms of the interruptions applied. Regarding Time to Perform Task, in-person interruption was found to have a significant effect, with the cost of a longer Time to Perform Task represented by the time taken to reorient to the task performance. A larger number of errors were found during task performance if either an in-person or instant messaging interruption took place.

The findings also showed that task load was significantly rated negatively during an interrupted performance. However, the Mental Load, Performance and Effort were only rated negatively if the interruption was carried out in person or as an instant message. Phone interruptions did not influence these measurements significantly compared to when there was no interruption. For the other measurements, Stress, Time Pressure and Frustration were rated negatively if any kind of interruption took place.

These results indicate that in-person interruptions are the most disruptive as they influence the number of errors, task load measurements and Time to Perform Tasks. Instant Messaging also influenced the number of errors (more errors) and Task Load measurements. Phone interruptions had little influence on the Performance measurements. Phone interruption was only rated significantly for some measurements of the Task Load, including Stress, Time Pressure and Frustration.

Table 6.11: Difference Between Work-load Measures Across Interruptions Along Significant Post-hoc Bonferroni Pair-Wise Comparisons Across Interruptions Scale Is 1 (Low) – 20 (High)

	Statistical Test	***p*-value**	**Effect Size**	**Statistical Power**	**Significant results of post-hoc Bonferroni analyses**
Mental load	F (2.589, 121.6) = 101	$p < 0.001$	d = 0.6, large	1- β = 1.00, perfect	B vs IM ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
					Ph vs Pr ($p = 0.021$)
Stress	F (2.22, 104.67) = 33.621	$p < 0.001$	d = 0.8, large	1- β = 1.00, perfect	B vs Ph, ($p < 0.001$)
					B vs IM ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
					Ph vs Pr ($p = 0.021$)
Time pressure	F (2.43, 114.64) = 25.92	$p < 0.001$	d = 0.7, large	1- β = 1.00, perfect	B vs Ph ($p < 0.001$)
					B vs IM ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
					IM vs Pr ($p = 0.015$)
Performance	F (2.22, 104.66) = 8.503	$p < 0.001$	d = 0.4, large	1- β = 0.9, very high	B vs IM, ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
Effort	F (2.59, 121.98) = 15.41	$p < 0.001$	d = 0.4, large	1- β = 0.9, very high	B vs IM ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
Frustration	F (2.49, 117.09) = 29.054	$p < 0.001$	d = 0.7, large	1- β = 1.00, perfect	B vs Ph ($p < 0.001$)
					B vs IM ($p < 0.001$)
					B vs Pr ($p < 0.001$)
					Ph vs IM ($p < 0.001$)
					Ph vs Pr ($p = 0.021$)

Table 6.12: Integration of Qualitative Data Indicated that IM is More Disruptive with Related Quantitative Data and Statistical Results

IM More Disruptive			
Qualitative Data	**Quantitative Results**		
'I am not quite sure about my performance for that task when you text me, though!'	Errors	IM vs B ($p < 0.001$) IM vs Ph ($p < 0.001$)	IM interruptions caused more errors in the task performance compared to no interruptions and phone interruptions.
'I found it more disruptive to handle the text messages, as every time I thought I would not get a new message and I was about to resume the task, a new message came'. 'It was hard to shift my mind between the task and the messages'. 'I found when you messaged me on WhatsApp it was really annoying because I was trying to focus on the tasks and answer you at the same time!' 'I was trying to get the task right and I was so focused on the task, but at the same time the messages got my mind away, really!'	Mental Load	IM vs B ($p < 0.001$) IM vs Ph ($p < 0.001$)	Following IM interruptions, participants rated the mental load higher compared to no interruptions and phone interruptions.
'It is quite stressful to answer the messages and resume the task'. 'To answer your messages… that was stressful'.	Stress	B vs IM ($p < 0.001$) Ph vs IM ($p < 0.001$)	IM interruptions caused participants to experience higher stress compared to no interruption and phone interruptions.
'But I couldn't do that. The messages were so frequently sent'. 'I was under pressure to reply to the messages as I wanted to return to the task as quickly as I could'.	Time pressure	IM vs B ($p < 0.001$) IM vs Pr ($p = 0.015$)	Following IM interruptions, participants rates time pressure higher compared to no interruption and in-person interruptions.
'I am not quite sure about my performance for that task when you text me, though!'.	Performance	IM vs B ($p < 0.001$) IM vs Ph ($p < 0.001$)	Following IM interruptions, participants rated their performance lower than when there is no interruption and following phone interruptions.
'I tried hard to shuffle between the messages and the task….'.	Effort	IM vs B ($p < 0.001$) IM vs Ph ($p < 0.001$)	Following IM interruptions, participants rated their effort higher than when there is no interruption, and following phone interruptions.
'I felt frustrated trying to answer your messages…'.	Frustration	IM vs B ($p < 0.001$) IM vs Ph ($p < 0.001$)	Following IM interruptions, participants rated their frustration higher than when there is no interruption and following phone interruptions.

Table 6.13: Integration of Qualitative Data Indicated that Pr is More Disruptive with Related Quantitative Data and Statistical Results

Pr More Disruptive			
Qualitative Data	**Quantitative Results**		
'I had more difficulties concentrating on the task after you popped in and start asking me questions'. 'I took me a bit of time to remember what I was specifically doing after you left'. 'That task when you asked me here took forever for me to find what it asked about'. 'I hardly remember what I had to do or how to complete the task'. 'That was… I stopped for a little bit before I continued working on the task'.	Time	Pr vs B ($p < 0.05$)	In-person interruptions caused longer actual task performance time compared to when there is no interruption.
'I am not sure whether I solved that task correctly when you came in'. 'I hope I have answered that task correctly when you asked me here in the room'.	Errors	Pr vs B ($p < 0.001$) Pr vs Ph ($p < 0.001$)	In-person interruptions caused participants to make more errors in the task performance compared to when there is no interruption and when phone interruptions were applied.
'When you came in and asked me, that was really disturbing for me'.	Mental Load	Pr vs B ($p < 0.001$) Pr vs Ph ($p = 0.021$)	In-person interruptions caused participants to rate the mental load higher than when there is no interruption and when phone interruptions are applied.
'It is quite stressful to be focusing on something and unexpected something happen like when you came in!'. 'I felt nervous when you suddenly came in'. .	Stress	Pr vs B ($p < 0.001$) Pr vs Ph ($p = 0.021$)	In-person interruptions caused participants to rate stress higher than when there is no interruption and when phone interruptions are applied.
	Time pressure	Pr vs B ($p < 0.001$) Pr vs IM ($p = 0.015$)	In-person interruptions caused participants to rate time pressure higher than when there is no interruption and when in-person interruptions are applied.
'I tried hardly to solve that task after you came here and talked to me, I think it was the hardest'.	Performance	Pr vs B ($p < 0.001$)	In-person interruptions cause participants to rate their performance lower than when there are no interruptions.
'I tried hardly to solve that task after you came here and talked to me, I think it was the hardest'.	Effort	Pr vs B ($p < 0.001$)	In-person interruptions cause participants to rate their effort lower than when there are no interruptions.
'To be honest I was a bit intimidated once you suddenly came in!'. 'That was frustrating to answer the question in front of you'.	Frustration	Pr vs B ($p < 0.001$)	In-person interruptions cause participants to rate their frustration higher than when there are no interruptions.

The qualitative findings support the quantitative results, as they showed that in-person interruptions were the most frequently mentioned cause of disruption during the testing, followed by instant messaging. The integration between the qualitative and quantitative findings highlighted this finding.

If we consider the significant cost in usability testing with in-person interruptions, we refer to longer Time to Perform Tasks, a higher number of errors, a higher Mental Load, more Effort and worse Performance. These also apply to the cost of instant messaging during the usability testing, except that Time to Perform Task did not lengthen significantly. If a phone call were received during the usability testing session, the participant felt pressure on their time, stressed and/or frustrated, but it was unlikely to lengthen the time taken to perform the task or make them commit more errors. Note that the increased time does not include the interruption itself.

Chapter 7: Discussions

7.1 Overview

This thesis has investigated the implication of applying usability testing with remote users in their natural environment. The findings support the assertion that for usability testing, what happens during a test session determines the quality and validity of data on users' performance. The usability testing method when applied and administered using online means and tools, such that it automatically records data on users' performance metrics and collects their subjective feedback, is independent of the testing environment.

This chapter highlights the relevant observations that can be drawn from the previous three chapters comprising this research presented in Chapters 4, 5, and 6, discusses their interpretations with relation to usability practise.

In this discussion section, the researcher intends to link the results from all prior research conducted in the field of RAUT and provide additional knowledge to the existing literature. A set of practical implications and recommendations for the usability practise community based on the observations and lessons experienced throughout this research will be provided.

7.2 Discussion of Key Findings

The present research provides a more holistic view than what is currently available in the literature that will extend our understanding of the implication of using usability testing with remote users, particularly RAUT, using online communication means. This holistic view is achieved by using empirical exploratory, explanatory comparative, and validation experimental research approaches conducted systematically and sequentially.

Unlike previous studies on RAUT, the exploratory empirical study derives important insights and lessons from representative participants in two kinds of representative environments (Lab and NE), investigating the usability testing outcomes, data, and reported feedback.

In the first two studies—exploratory and explanatory—participants performed identical experimental usability testing tasks in two different environments (Lab and NE), where NE group served as the control group, enabling valid comparison between their performances

interacting with real digital library websites and behaving as if they were performing real searching and retrieving tasks. The third controlled experimental study, on the other hand, simulated interruptions based on the applicable influential participants' contextual reports from the first two empirical studies regarding distractions. This design enabled the factor of interest - the interruption, based on the first two studies' suggested findings - to be isolated and investigated for its influence on usability testing outcomes, and produced new knowledge regarding the implication of RAUT with users in their ordinary natural environment to usability practise. The following sections will discuss the key results associated with the literature and related works.

7.2.1 Contextual Factors and Usability Testing

Typically, a RAUT method takes place with participants in their natural environment to gain insight into the actual realistic users' interaction with the evaluated system. Specifically, RAUT is based on online un-moderated communication to understand the level to which we can adopt and trust the data on user performance during usability testing in the participant's natural environment. The first and second studies in this research were conducted to ascertain and understand data trustworthiness in RAUT.

Both studies showed that usability practitioners should consider the so-called 'completion time' in the literature with caution. In both studies, the completion time was found to reflect different meanings besides the actual performance time on the tasks. As discussed in Chapter 2, several studies have referred to completion time as the time to perform the test tasks; yet, the setup of the test applied within those studies incorporates the whole time consumed during the test in this measurement. In such a situation, it might be more accurate to call it 'time to complete the test'. Interestingly, the time it takes to complete a test is a factor used in psychology, also measures the level of distraction. We can now see how risky it is to consider completion time this way to represent the Time to Perform Task(s). Completion time is therefore only a tool to demonstrate distraction, and the Time to Perform Task is meant to measure the exact time consumed during task performance only.

7.2.2 RAUT Evaluation Method and The Type of Environment

This section will focus on the differences between the two environments regarding usability testing outcomes. The first and second studies have shown that usability testing outcomes

are independent of the RAUT method itself. The justification of this statement is that both studies yielded similar results when usability testing was applied in the two different environments; the difference was related to contextual factors. Therefore, whether the test was conducted in a lab or not is insignificant compared to what happens during the test itself. This was also evident in the third study, when differences were found in the participants' performance between interrupted tasks and non-interrupted tasks, and all these tasks were carried out in the same lab environment.

7.2.3 The Cost of Interrupted Performance in Usability Testing

The cost here is represented by how much the actual performance would differ from if there is no distraction. This is translated as the time required to reorient to the tasks which will ultimately lengthen the Time to Perform Task, in addition to the increased number of errors.

Based on the third study's findings, the interruption can have a negative influence on "cost" on usability testing outcomes, yet the extent of this influence is also different based on the type of interruption applied. Time to Perform Task is only significantly lengthened by the in-person interruption. Instant messaging significantly increases the number of errors. One possible interpretation is that when instant messaging takes place, the participant might shuffle between the two platforms—the machine where the test is running and the phone—in this way the Time Per Task would not be influenced is no need to reorient to the task as the participant is still performing*, for example, referring to Table 6.12, one participant said, "I found it more disruptive to handle the text messages, as every time I thought I will not get a new message and I am about to resume the task a new message come". While for the in-person interruption, participants have explicated more frustrated feedback, referring to Table 6.13, one participant said, "I had more difficulties concentrating on the task after you popped in and start asking me". This total mental focus shift and frustration might take participants a few minutes to re-concentrate again and reorient to the task.

We can see that in-person interruption and instant messaging significantly increased the Task-Load in terms of time, mental load, and effort. Phone interruption has been found to have a negative influence in terms of the ratings due to time pressure, stress, and frustration.

* Although they have been informed not to do so, see Appendix A.CH6.1: Information Sheet

7.3 Discussion Notes

Few have investigated the influence of different testing environments on usability testing outcomes. For example, Andrzejczak and Liu (2010) investigated the effect of test location (lab vs. remote) on usability testing performance, participant stress level, and subjective testing experience. They adopted UCI reports in the remote setting, and the test was applied synchronously.

Khanum and Trivedi (2013) investigated the effects of the testing environment on usability testing outcomes using TAP with children in the unfamiliar lab room and a familiar computer lab (field setting), an approach similar to the local remote testing described by Hartson et al. (1996).

Both studies remarked on the high possibility of the distractions' presence in the remote/field environment. Andrzejczak and Liu (2010) stated, "Distractions and stressors may be present and not controlled in the remote laboratory setting such as disruptive students, fire drills, and other distractions present in a high-traffic environment" (p. 1265) while Khanum and Trivedi (2013) stated that "In the field test, there were interruptions as no restrictions were imposed on the people to move in the field, but these did not affect the performance much" (p. 2052). Both studies have not attempted to gather data about these distractions to relate the differences found, if any, to them.

Greifeneder's (2011) study was conducted in both settings: lab and remote, which was applied and administered online. Her study gathered data about distractions during the natural environment session, and she attempted to investigate whether there is a relationship between the distractions reported and differences found.

The agreement of this research with previous studies findings or interpretations can be summarised as participants consuming a longer time performing the test in a natural environment than in the lab environment. For example, Greifeneder's (2011) findings stated that "people in the natural environment needed statistically more time to complete the test" (p. 312). Yet, one could not conclude whether the few differences found were *due* to the contextual factors reported by participants in the remote setting. Our research has further shown that contextual factors such as interruptions and connection speed will influence the whole time required to perform the test, yet participants, whenever they can, will not allow these interruptions during task performance.

Our research collected data from participants regarding distractions (if they occur) during their testing session. We have simulated interruptions like the workflow study of Mark et al. (2008), which concluded that interrupted participants work faster but at a price—higher workload, higher frustration, more stress, more time pressure, and effort. They tried to interpret these phenomena and stated that "another possibility is that interruptions do lengthen the time to perform a task but that this extra time only occurs directly after the interruption when reorienting back to the task, and it can be compensated for by a faster and more stressful working style" (p. 110). Our results showed that interruption leads participants to consume a longer time performing the task, but only if it was by in-person interruption where the subjective workload in terms of performance, effort, and where mental load has a higher negative rating. The participants' feedback also stressed that interruption by a person was frustrating and caused higher shifting in their mental state. For other interruptions, such as instant messaging and phone, the participants also consumed more time performing the task, but it was not as significant. The findings of the present research and Mark et al. (2008) are different; however, we should not forget that the context of the two experiments was different: information workflow and usability testing, hence the tasks given for performance in the two experiments were different. With usability testing tasks, participants might feel less guilty if they were exposed to interruption and may feel they have the right to take time to re-concentrate on the task. Alternatively, participants might feel they should find the answer to the tasks as they were problem-solving tasks, so they would not try to compensate for the time elapsed on the interruption by working faster after the interruption.

To gain more insight, the methodological research investigated three studies in order to explore the distractions that occur during usability testing, address them, control for the differences in the data collection method, control or account for the confounding variables, and then, use the insights and findings to investigate how these distractions influence the usability testing outcomes, controlling for all the confounding variables while achieving both validities: external (study 1 and 2) and internal (study 3). Hence, this research was able to show the trends of the differences in usability testing outcomes, correlation and the amount of variance in the usability testing outcomes, and finally, the source of influence on the usability testing outcomes (see Table 7.1).

7.4 Implication of Applying Usability Testing with Remote Users

As discussed in Chapter 1, section 1.2, the objective of studying RAUT is to maximize its benefits and comprehend its shortcomings to get the most out of the testing data provided by it rather than just comparing it with other evaluation methods.

Based on our research, we stressed that if RAUT is the usability evaluation method used, then the usability practitioner and researcher should expect some contextual factors that would influence the data to be collected out of the method. Therefore, we recommend that:

- For the usability practitioner:
  - The clarity of the language of the textual descriptions used in the usability testing transcripts is very important, and the language should be appropriately used according to the level of participants.
  - Task(s) start, and end should be designed to be highly noticed by participants. If more than one task is to be performed, they should be named accordingly, as this would make it easier for the participants to realise which task was interrupted.
  - Time measurements should be collected in two variables: Time on Tasks, which is solely reflecting the time consumed performing the tasks only. Other times should be represented by another variable, for example, Time on Questions.
  - Contextual factors must be addressed either by:
    - Experimentally controlling for them: video or/and audio recording using some recent unmoderated online usability testing tools.
    - Statistically mitigating their influence by dealing with outlier values and validating the results with post-interview aimed to know what was happening (e.g., what interruption(s) happened, and the apparatus used)
- For the usability testing research and/or technological development community:
  - A great innovation would be to develop or enhance the unmoderated tool that detects the interruption triggers or signs and produce a timeline report for all instances that happen during the test. For example, recording video, audio, and screen of the participants can detect if the curser was idle and tracking to see if the eyes were not toward the screen or if the participant's voice was on for more than 5 seconds or so.

Table 7.1: Filling the Gap in This Research

		Was the study an Empirical Comparative study?	Were the same data collection methods used?	Type of environments	Adopted usability testing for the non-lab environment		Participants		Were contextual factors gathered?	Was the Relationship between the outcomes and the contextual factor investigated?	Is there a Conclusion about the source of the difference?	Validity
					Asynchronous or synchronous?	Formative or summative?	**Type**	**Sample size**				
Andrzejczak and Liu (2010)		√	×	Lab and Field (both adopted in the university lab rooms)	Synchronous	Both	Adults	60 (30:30)	×	×	×	External (questionable)
Khanum and Trivedi (2010)		√	√	Lab and Field (both adopted in the school rooms)	Asynchronous	Formative	Children	18 (9:9)	×	×	×	External
Greifeneder (2011)		√	Yes	Lab and NE	Asynchronous	Summative	Adults	31 (13:18)	√	√	×	External
The research methodological approach for this thesis	**Study 1**	√	Yes	Asynchronous	Lab and NE	Predominate Summative + Formative	Adults	30 (10:20)	√	×	Tendencies	External
	Study 2	√	Yes	Asynchronous	Lab and NE	Predominate Summative + Formative	Adults	96 (48:48)	√	√	Correlation & Amount of variance	External
	Study 3	√	Yes	Asynchronous	Lab and NE	Summative	Adults	48	√	√	√	Internal

Chapter 8: Conclusions

8.1 Overview

This final chapter draws out the conclusions of the research. It starts by summarising the research and its major findings, and then moves on to evaluate whether the aims and objectives of the research were achieved. This is followed by a section identifying the key contributions that have been made to the body of knowledge. After a discussion of the limitations of the research, the chapter concludes by suggesting potential avenues for future work.

8.2 Evaluation of Research Aim and Objectives

After developing a background context for the research, the research motivations were defined, from which the research aim, and objectives were drawn. As discussed in the first chapter, this research has been undertaken through a series of empirical studies using formal empirical summative online usability studies to achieve the research aims. This research achieved the following objectives:

- Exploring the functionality of usability studies, in administering the test, and its tasks, instructions, and questions within different experimental settings.
- Exploring the data provided by participants through the online administrated usability study about the interaction with the test object(s) during the testing session in the different testing environments.
- Exploring usability outcomes in different testing environment settings.

That was achievable by the exploratory study, Chapter 4.

- Investigating the contextual factors reported by remote participants during their RAUT session.
- Investigating the difference in usability testing outcomes, in terms of participants' performance and subjective ratings, in different testing environment settings.
- Investigating the relationship between the contextual factors reported by participants and the differences in the usability testing outcomes, if any.

That was achievable by the explanatory comparative study, Chapter 5.

- Validating the findings from the exploratory and explanatory studies in terms of relationship found between the interruptions and time measurements.
- Isolating the possible source of effect, interruption, applying an experiment in which all confounding variables would be controlled in to investigate its effect on usability testing outcomes.
- Investigating the differences in usability testing performance between the interrupted tasks and the non-interrupted task performance.
- Investigating the differences between the task-load incurred by the interrupted tasks and the non-interrupted task performance.
- Investigating the interruption cost in term of how the task(s) performance would be influenced by interruptions.
- Understanding which type of interruption is the most disruptive for participants to perform the task.
- Understanding why participants perform tasks poorly when interrupted and by which interruption with their own feedbacks and opinions. type of interruption is the most disruptive for participants to perform the task.

That was achievable by the validation study, Chapter 6.

8.3 Novelty and Contribution to The Body of Knowledge

The novelty of this research and the key contributions are as follows:

- Analysing the literature extensively on RAUT and studies investigate factors on usability testing outcomes, Table 2.5.
- Mapping 4FFCF model with theories relevant to distractions and its influence, Table 3.5. extensively on RAUT and studies investigate factors on usability testing outcomes, Table 3.2.
- Mitigating the validity issues acknowledged in the relevant literature in terms of the limitations in the statistical tests applied, the conclusions passed to the practitioners and researchers, and the instrumentations and measures used for comparison(s).
- To the best of the author's knowledge, this is the first contribution that simulates interruptions and examines their effect on usability testing outcomes with the aim of understanding the implications for usability testing practise.
- Ensuring the reliability of comparisons, by avoiding the evaluator effect.

- Understanding the relationship between contextual factor implied in usability testing session and the usability testing data.
- Ensuring proper assessment of the usability testing practical utility by focusing on the design impact along with users' feedbacks on usability.
- Ensuring the validity of comparisons, instrumentations and generalisability of results for usability testing outcomes in different environments,
- Applying a new approach to investigate the capabilities and shortcomings of usability testing with remote users.
- Considering contextual factors and their possible impact on usability testing.
- Ensuring the awareness of the possibility of the existence of contextual factors, their types and frequency, and considering their possible relationships to usability testing outcomes.
- Exploring and investigating the source of the inconsistencies in the results reported by RAUT's outcomes in the literature compared to traditional usability testing.
- Drawing results as to whether the inconsistencies in the results reported by RAUT's outcomes in the literature compared to the traditional usability testing were related to the usability testing methods used or to the testing environment utilised.
- Investigating the implication of the existence of influential contextual factors during usability testing performance on its outcomes.
- Investigating the implication of the existence of influential contextual factors during usability testing performance on its outcomes.
- Filling the gap of the experimental validation for the *physical* testing environment factor of the 4FFCF model, which was proposed by Sauer et al, (2010) on usability testing outcomes. The sources of the influence are shown in Table 8.1 along with the influenced usability testing outcomes.

Table 8.1: Experimentally Validated Influential Physical Environment's Factors on Usability Testing Outcomes

Theory				Mapping to 4FFCF model		
				Condition(s)/Facto(s) Environment		Expected implication(s) on Usability testing outcomes
Name	**Conditions**	**Source of influence**	**Theory suggested implications**	**Distraction type**	**Source**	
Social facilitation theory	Presence of others × task complexity	Amplified arousal	• Improved performance of an easy task **(**Fraser et al., 2001). • No change in the performance of easy tasks ((Baron, 1986); (Manstead & Semin, 1980)). • Deteriorate performance for complex **((**Fraser et al., 2001), (Baron, 1986), & Manstead & Semin, 1980)).	External interruption	**In-person conversation**	Performance • **Time on Tasks** • Successful completions • Number of page views • **Errors**
Distraction-conflict theory	Distractions × task complexity	Cognitive overload	• Concentration on a small number of cues lead to improved quicker performance of an easy task (Baron, 1986). • Attention is required to be paid to a stimulus of complex task while handling the information presented from the distracting task. (Bernd, 2002). • Change in complex task processing (March, 1994). • Reduced performance accuracy of complex task (Cellier and Eyrolle, 1992). • Change in information use from complex task (Baron, 1986). • Longer time to solve complex task (Cohen, 1980; Malhotra et al., 1982).	External interruption	• **In-person conversation** • Phone calls • **Intrusive text messages**	Performance • **Time on Tasks** • Successful completions • Number of page views • **Errors**
Information-overload	Information cues, task demand, or task complexity	Limited cognitive processing capacity (Mental workload)	• Reduction in the quality of decisions made ((Speier et al., 1999; (Chewning and Harrell, 1990; Snowball, 1980)). • Increasing the time needed to make a decision (Cohen, 1980; Malhotra et al., 1982). • Misunderstandings and confusions concerning the decision (Cohen, 1980; Malhotra et al., 1982).	External interruption	• **In-person conversation** • Phone calls • **Intrusive text messages**	Performance • **Time on Tasks** • Successful completions • Number of page views • **Errors** Perceived usability • Ratings • Reports
				Multitasking	Other opened applications	
				Poor apparatus performance	• Small display size • Low connection speed	

References

Adamczyk, P. D., and Bailey, B. P. (2004). If not now, when? The effects of interruption at different moments within task execution. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 271–278.

Äijö, R., and Mantere, J. (2001). Are non-expert usability evaluations valuable? *18th International Symposium on Human Factors in Telecommunications*, Bergen, Norway, 5-7 November 2001, viewed 30 March 2014, <http://www.hft.org/HFT01/paper01/acceptance/2_01.pdf>.

Anandhan, A., Dhandapani, S., Reza, H., and Namasivayam, K. (2006). Web Usability Testing—CARE Methodology. *Third International Conference on Information Technology: New Generations (ITNG'06)* IEEE, pp. 495-500.

Albert, B., Tullis, T., and Tedesco, D. (2009). Beyond the usability lab: Conducting large-scale online user experience studies. Burlington, MA: Morgan Kaufmann.

Allport, F. H. (1920). The influence of the group upon association and thought. *Journal of Experimental Psychology*, 3(3), p. 159.

Altmann, E. M., and Trafton, J. G. (2004). Task interruption: Resumption lag and the role of cues. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26(26).

Altmann, E. M., and Trafton, J. G. (2007). Time course of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review*, 14(6), pp. 1079–1084.

Andreasen, S., Nielsen, V., Schrøder, O., and Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, USA, pp. 1405–1414.

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

Atchley, P., and Chan, M. (2011). Potential benefits and costs of concurrent task engagement to maintain vigilance: A driving simulator investigation. *Human factors*, 53(1), p. 312.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 6, pp. 574–594.

Baron, R. S. (1986). Distraction-conflict theory: Progress and problems. *Advances in Experimental Social Psychology*, 19, pp. 1–39.

Batra, S., and Bishu, R. (2007). Web usability and evaluation: Issues and concerns. N. Aykin (ed.). Usability and Internationalization. *Proceedings of the HCII 2007 conference*. Berlin, Heidelberg: Springer-Verlag, 4559, pp. 243–249.

Baym, N. K., and Markham, A. N. (2009). Introduction: Making smart choices on shifting ground. A. N. Markham and N. K. Baym (Eds.), *Internet inquiry: Conversations about method*, Sage Publications, Inc., pp. vii–xix.

Beeftink, F., Van Eerde, W., and Rutte, C. G. (2008). The effect of interruptions and breaks on insight and impasses: Do you need a break right now?. *Creativity Research Journal*, 20(4), pp. 358–364.

Bell, E., Bryman, A., and Harley, B. (2018). *Business research methods*. Oxford university press.

Bernd, S. (2002). Social facilitation in motor tasks: A review of research and theory. *Psychology of Sport and Exercise*, 3(3), pp. 237–256.

Bevan, N., and Macleod, M. (1994). "Usability Measurement in Context", *Behaviour and Information Technology (BIT)*, 13(1-2), pp. 132–145.

Bond, C. F. (1982). Social facilitation: A self-presentational view. *Journal of Personality and Social Psychology*, 42(6), p. 1042.

Borgholm, T., and Madsen, K.H., (1999). Cooperative usability practices. *Communications of the ACM*, 42 (5), pp. 91–97.

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think-alouds and think afters. *Library and Information Science Research*, 22(4), pp. 371–392.

Brewer, M. B., and Crano, W. D. (2000). Research design and issues of validity. *Handbook of Research Methods in Social and Personality Psychology*, Cambridge University Press, pp. 3–16.

Brixey, J.J., Robinson, D. J, Johnson, C.W., Johnson, T.R., Turley, J.P., and Zhang, J. (2007). A Concept Analysis of the Phenomenon Interruption, *Advances in Nursing Science*. 30(1), E26.

Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. P. W. Jordan, B. Thomas, B.A. Weerdmeester, and I.L. McClelland (eds.), *Usability Evaluation in Industry*, London: Taylor & Francis, pp. 189–194.

Brumby, D. P., Cox, A. L., Back, J., and Gould, S. J. (2013). Recovering from an interruption: Investigating speed-accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, 19(2), p. 95.

Brush, J., Ames, M., and Davis, J. (2004). A comparison of synchronous remote and local usability studies for an expert interface. CHI EA 2004: *Proceedings of the CHI 2004 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, USA, pp. 1179–1182.

Bruun, A., Gull, P., Hofmeister, L., and Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. *CHI 2009: Proceedings of the SIGCH Conference on Human Factors in Computing Systems*, Boston, USA, 4-9 April 2009. ACM, New York, USA, pp. 1619–1628.

Bryman, A. (2003). *Quantity and quality in social research*, 18. Routledge.

Bryman, A., and Bell, E. (2011). *Business research methods.* Oxford Press, Oxford.

Buchanan, E. A. (2004). Readings in virtual research ethics: Issues and controversies. Hershey, PA: Information Science Publishing, p. 362.

Campbell, D. J. (1988). Task Complexity: A Review and Analysis, *Academy of Management Review*, 13(1), pp. 40–52.

Card, S. K., Moran, T. P., and Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Erlbaum.

Castillo, J. C., Hartson, H. R., and Hix, D. (1998). Remote usability evaluation: Can users report their own critical incidents? *CHI 98: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Los Angeles, USA, 18-23 April 1998. ACM, New York, USA, pp. 253–254.

Cellier, J., and Eyrolle, H. (1992). Interference between switched tasks. *Ergonomics*, 35(1), pp. 25–36.

Chewning, E. G., Jr., and Harrell, A. M. (1990). The effect of information load on decision-makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations, and Society*, 15(6), pp. 527–542.

Chisholm, C. D., Dornfeld, A. M., Nelson, D. R., and Cordell, W. H. (2001). Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices. *Annals of emergency medicine*, 38(2), pp. 146–151.

Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: A review of research and theory. *Psychological Bulletin*, 88(1), p. 82.

Colligan, L., and Bass, E. J. (2012). Interruption handling strategies during paediatric medication administration. *BMJ Qual Saf*, 21(11), pp. 912–917.

Covey, S. R. (1989). *The 7 Habits of Highly Effective People*. New York: Simon and Schuster.

Creswell, J. W. (2003). *Research Design: Qualitative. Quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage.

Creswell, J. W. (2013). *Research Design. International Student Edition*. SAGE Publications Ltd.

Creswell, J. W., and Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage Publications.

Creswell, J. W., and Plano Clark, V. L. (2017). *Designing and conducting mixed methods research*. SAGE Publications Ltd.

Croskerry, P. (2013). From mindless to mindful practice—cognitive bias and clinical decision making. *N Engl J Med*, 368(26), pp. 2445–2448.

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Sage.

Czerwinski, M., Horvitz, E., and Wilhite, S. (2004). A diary study of task switching and interruptions. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 175-182.

Denscombe, M. (2003). *The Good Research Guide for Small-scale Social Research Projects*. Open University Press.

Denzin, N. K., and Lincoln, Y. S. (2011). *The Sage handbook of qualitative research*. SAGE.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, California: Sage.

Dix, A.Finlay, J., Abowd, G.D., and Beale, R. (2004). Human-Computer Interaction. Pearson Education, Ltd, Harlow, pp. 318–364.

Downing, C., and Liu, C. (2011). Assessing web site usability in retail electronic commerce. *IEEE 35th Annual Computer Software and Applications Conference (COMPSAC)*, pp. 144–151.

Dray, S., and Siegel, D. (2004). Remote possibilities? International usability testing at a distance. *Interactions*, 11(2), pp. 10–17.

Dumas, J. S., and Fox, J. E. (2008). Usability testing: Current practice and future directions. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (Second ed), New York: Erlbaum, pp. 1129–1149.

Elliot, L.R., Dalrymple, M.A., Schifflett, S.G., and Miller, J.C. (2004). Scaling scenarios: Development and application of C4ISR sustained operations research. Schiflett, S.G., Elliot, L.R., Salas, E., and Coovert, M.D. (eds.), *Scaled worlds: Development, validation, and applications.* Ashgate Publishing Limited, Aldershot, pp. 119–133.

Ericsson, A., and Simon, A. (1998). How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), pp. 178–186.

Ernest, P. (1994). *An Introduction to Research Methodology and Paradigms.* Exeter: School of Education, University of Exeter.

Evaristo, R., Adams, C., and Curley, S. (1995). Information load revisited: A theoretical model. *Proceedings of the 16th Annual International Conference on Information Systems*, Amsterdam, pp. 197–206.

Faulkner, X. (2000). *Usability Engineering.* London: Macmillan Press.

Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4(1), pp. 6–16.

Finstad, K. (2006). The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, 1, pp. 185–188.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies*, 5(3), pp. 104–110.

Fitzpatrick, R. (1998). Strategies for Evaluating Software Usability. *Methods*, 353(1). doi.org/10.21427/gxqn-pw69.

Flick, U. (1998). *An Introduction to Qualitative Research*, SAGE Publications Ltd.

Fraser, C., Burchell, B.J., Hay, D., and Duveen G. (2001). *Introducing social psychology.* Oxford: Polity, pp. 15–20.

Gable, G. G. (1994). Integrating case study and survey research methods: An example in information systems. *European Journal of Information Systems*, 3(2), pp. 112–126.

Gillie, T., and Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, *50*(4), pp. 243–250.

Goddard, W., and Melville, S. (2004). *Research Methodology: An Introduction* (2nd ed.) Oxford: Blackwell Publishing.

Godfrey, P., and Hill, C. (1995). The problem of unobservable in strategic management research. *Strategic Management Journal*, 16(5), pp. 19–533.

Gray, W. D., and Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13(3), pp. 203–261.

Greifeneder, E. (2011). The impact of distraction in natural environments on user experience research. *Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg, pp. 308–315.

Groff, B. D., Baron, R. S., and Moore, D. L. (1983). Distraction, attentional conflict, and drive like behavior. *Journal of Experimental Social Psychology*, 19(4), pp. 359–380.

Grundgeiger, T., and Sanderson, P. (2009). Interruptions in healthcare: Theoretical views. *International Journal of Medical Informatics*, 78(5), pp. 293–307.

Guba, E. G., and Lincoln, Y. S. (1994). Competing paradigms in qualitative research. Denzin, N.K., and Lincoln, Y.S. (Eds.), *Handbook of qualitative research*, Sage Publications, Inc., pp. 105–117.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., and Herulf, L. (2004). Making a difference: a survey of the usability profession in Sweden. *Proceedings of the third Nordic conference on Human-computer interaction—NordiCHI'04,* pp. 207–215.

Gupta, A., Li, H., and Sharda, R. (2013). Should I send this message? Understanding the impact of interruptions, social hierarchy and perceived task complexity on user performance and perceived workload. *Decision Support Systems*, 55(1), pp. 135–145.

Hart, S. G. (1986). Theory and measurement of human workload. J. Zeidner (eds.), *Human productivity enhancement: Training and human factors in system design*. New York: Praeger, pp. 396–456.

Hartson, H. R., and Castillo, C. (1998). Remote evaluation for post-deployment usability improvement. *Proceedings of the Working Conference on Advanced Visual Interfaces*, L'Aquila, Italy, 24-27 May 1998. ACM, New York, USA, pp. 22–29.

Hartson, H.R., Andre, T.S., and Williges, R.C. (2001). Criteria for Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), pp. 373–410.

Hartson, H. R., Castillo, J. C., Kelso, J., and Neale, W. C. (1996). Remote evaluation: The network as an extension of the usability laboratory. *CHI 1996: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, 13–18 April 1996. ACM, New York, USA, pp. 228–235.

Hertzum, M., and Jacobsen, N.E. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), pp. 421–443.

Hine, C. (2006). Virtual methods: Issues in social research on the internet. Oxford: Berg. Hilbert, M., and Redmiles, F. (1999). *Separating the Wheat from the Chaff in Internet-Mediated User Feedback Expectation-Driven Event Monitoring*. ACM SIGGROUP Bulletin, 20(1), pp. 35–40.

Hintze, J. M., Volpe, R. J., and Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. *Best practice in school psychology*, 4, pp. 993–1006.

Hirschheim, R. (1985). Information systems epistemology: A historical perspective. *Research methods in information systems*, pp. 13–35.

Hirschheim, R. (1992). Information Systems Epistemology: An historical perspective. *Information Systems Research. Issues, methods, and practical guidelines.* Oxford: Blackwell Scientific Publications, 39.

Hix, D., and Hartson, H. R. (1993). *Developing user interfaces: Ensuring usability through product and process.* New York, NY: John Wiley and Sons.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, 29(1), pp. 97–111.

IEEE. (1990). *IEEE standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries.* New York: IEEE.

ISO 9241-11 (1998). *Ergonomic requirements for office work with visual display terminals (VDTs). Part 11, Guidance on usability*, ISO.

ISO/IEC. (2001). *Software Engineering- Product Quality—Part 1: Quality Model.* Geneva, Switzerland: International Organisation for Standardisation.

Ivory, M. Y., and Hearst, M. A. (2001). The state-of-the-art in automated usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4), pp. 470–516.

Jaäskeläinen, R. (2001). Think-aloud protocols. *Routledge Encyclopaedia of Translation Studies*, pp. 269–273.

Jin, J., and Dabbish, L. A. (2009). Self-interruption on the computer: A typology of discretionary task interleaving. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1799–1808.

Kahneman, D. (1973). *Attention and effort*, 1063. Englewood Cliffs, NJ: Prentice-Hall.

Kaplan, B., and Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: A case study. *MIS quarterly*, pp. 571–586.

Katidioti, I., and Taatgen, N. A. (2014). Choice in multitasking: How delays in the primary task turn a rational into an irrational multitasker. *Human factors*, 56(4), *pp. 728–736.*

Kelly, D., and Gyllstrom, K. (2011). An examination of two delivery modes for interactive search system experiments: remote and laboratory. *Proceedings of the CHI '11 conference*, New York, NY: ACM. pp. 1531–1540.

Kessner, M., Wood, J., Dillon, R.F., and West, R.L., 2001. On the reliability of usability testing Conference on Human Factors in Computing Systems. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, pp. 97–98.

Khanum, A., and Trivedi, C. (2013). Comparison of Testing Environments with Children for Usability Problem Identification. *International Journal of Engineering and Technology*, 5(3), pp. 2048–2053.

Kirakowski, J., and Cierlik, B. (1998). Measuring the usability of web sites. *Proceedings of the Human Factors and Ergonomics Society annual meeting*, Sage CA: Los Angeles, CA: SAGE Publications, 42(4), pp. 424–428.

Laird, D. A., Laird, E. C. L., and Fruehling, R. T. (1983). *Psychology, human relations, and work adjustment*. Gregg Division, McGraw-Hill.

Lazar, J. (2006) Web Usability: A User-Centered Design Approach. Boston: Addison-Wesley.

Lazar, J., Feng, J. H., and Hochheiser, H. (2010). Research methods in human-computer interaction. John Wiley and Sons.

Lebiere, C., Anderson, J.R., and Bothell, D., 2001. Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task. *Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation.* Norfolk, VA, USA.

Lee, A. S. (1991). Integrating positivist and interpretive approaches to organizational research. *Organization Science*, 2(4), pp. 342–365.

Lee, B. C., and Duffy, V. G. (2015). The effects of task interruption on human performance: A study of the systematic classification of human behavior and interruption frequency. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 25(2), pp. 137–152.

Leedy, P. D., and Ormrod, J. E. (2005). *Practical research*. Pearson Custom.

Lewis, C., and Rieman, J. (1994). Task-centred User Interface Design. Available: ftp://ftp.cs.colorado.edu Accessed on 22/09/2016.

Lewis, J. (2006). Usability testing. Salvendy, G. (ed.), *Handbook of Human Factors and Ergonomics*. New York: John Wiley, pp. 1275–1316.

Lewis, J. R., and Sauro, J. (2009). The factor structure of the System Usability Scale. M. Kurosu (ed.), *Human-Centered Design*, HCII 2009. Berlin, Germany: Springer-Verlag, pp. 94–103.

Li, S. Y., Blandford, A., Cairns, P., and Young, R. M. (2008). The effect of interruptions on post completion and other procedural errors: An account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied*, 14(4), p. 314.

Li, S. Y., Magrabi, F., and Coiera, E. (2012). A systematic review of the psychological literature on interruption and its patient safety implications. *Journal of the American Medical Informatics Association*, 19(1), pp. 6–12.

Lincoln, Y. S., Lynham, S. A., and Guba, E. G. (2011). Paradigmatic controversies, contradictions, and emerging confluences, revisited. *The Sage Handbook of Qualitative Research*, 4, pp. 97–128.

Lindgaard G. and Chattratichart J. (2007). "Usability testing: What have we overlooked?", *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI'07*, ACM, April, San Jose, U.S.A., pp. 1415–1424.

Liu, D., Grundgeiger, T., Sanderson, P. M., Jenkins, S. A., and Leane, T. A. (2009). Interruptions and blood transfusion checks: Lessons from the simulated operating room. *Anesthesia and Analgesia*, 108(1), pp. 219–222.

Loukopoulos, L. D., Dismukes, R. K., and Barshi, (2009). The perils of multitasking. *AeroSafety World*, 4(8), pp. 18–23.

MacCallum, R.C., Zhang, S., Preacher, K.J. and Rucker, D.D., (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), p.19.

MacKenzie, I. S. (2013). *Human-computer interaction: An empirical research perspective.* Waltham, MA: Elsevier Morgan Kaufmann.

Maguire, M. (2001). Context of Use within Usability Activities. *International Journal of Human-Computer Studies*, Taylor and Francis Group, 55(4), pp. 453–483.

Malhotra, N. K., Jain, A. K., and Lagakos, S. W. (1982). The information overload controversy: An alternative viewpoint. *Journal of Marketing*, 46(2), pp. 27–37.

Manstead, A. S. R., and Semin, G. R. (1980). Social facilitation effects: Mere enhancement of dominant responses?. *British Journal of Social and Clinical Psychology*, 19(2), pp. 119–135.

March, J. G. (1994). *A primer on decision making: How decisions happen.* New York: The Free Press.

Mark, G., Gudith, D., and Klocke, U. (2008). The cost of interrupted work: More speed and stress. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 107–110.

Mark, G., Voida, S., and Cardello, A. (2012). "A pace not dictated by electrons" an empirical study of work without email. *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 555–564.

Marshall, C., and Rossman, G. (1999). *Designing qualitative research.* Newbury Park, CA: Sage.

May, T. (2011). *Social research: Issues, methods and research* (4th ed.). London: McGraw-Hill International.

Mayers, A. (2013). *Introduction to Statistics and SPSS in Psychology*. Harlow: Pearson Education.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., and Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC medical research methodology*, 7(1), p. 30.

McCoy, J.M., and Evans, G.W. (2005). Physical work environment. Barling, J., Kelloway, E.K., and Frone, M.R. (eds.), *Handbook of work stress.* London: Sage, pp. 219–245.

McFarlane, D. C. (1997). *Interruption of people in human-computer interaction: A general unifying definition of human interruption and taxonomy*. Arlington, VA: Office of Naval Research.

McFarlane, D. C., and Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, 17(1), p. 161.

Mertens, D. M. (2010). Philosophy in mixed methods teaching: The transformative paradigm as illustration. *International Journal of Multiple Research Approaches*, 4(1), pp. 9–18.

Millen, D. R. (1999). Remote usability evaluation: User participation in the design of a web-based email service. *ACM SIGGROUP Bulletin*, 20(1), pp. 40–45.

Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B., 2004. Comparative usability evaluation. *Behaviour & Information Technology*, 23, pp. 65–74.

Monk, C. A., Trafton, J. G., and Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, 14(4), p. 299.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of mixed methods research*, 1(1), pp. 48–76.

Murphy, E., Malakhoff, L., and Coon, D. (2007) Evaluating the usability and accessibility of an online form for Census Data Collection. Lazar, J. (ed.). *Universal Usability*. Chichester: John Wiley & Sons Ltd, pp. 517–558.

Neuman, W. L. (2009). *Understanding research*. Pearson.

Newman, I., Benz, C. R., and Ridenour, C. S. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum*. SIU Press.

Newsom, J., (2019, September 8) USP 634 data analysis. *Levels of measurement and choosing the correct statistical test.* Retrieved from http://web.pdx.edu/~newsomj/uvclass/

Nichols, A. L., and Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, 135(2), pp. 151–166.

Nielsen, J. (1990). Paper versus computer implementations as mockup scenarios for heuristic evaluation. *Proceedings of the IFIP TC13 Third International Conference on Human-Computer Interaction*, pp. 315–320.

Nielsen, J. (1993). *Usability engineering.* Boston: Academic Press, pp. 373–380.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, Ind.: New Riders.

Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people—CHI'90.* pp. 249–256.

Nilsson, J., and Siponen, J. (2005). Challenging the HCI concept of fidelity by positioning Ozlab prototypes. *Proceedings of the Fourteenth International Conference on Information Systems Development*, pp. 349–360.

Nivala, A.-M., and Sarjakoski, L. T. (2003). Need for Context-Aware Topographic Maps in Mobile Devices. *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science* (ScanGIS'03), pp.15–29.

Norman, D. A., and Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive psychology*, 7(1), pp. 44–64.

Nugus, P., and Braithwaite, J. (2010). The dynamic interaction of quality and efficiency in the emergency department: Squaring the circle?. *Social Science & Medicine*, 70(4), pp. 511–517.

Orgad, S. (2009). How Can Researchers Make Sense of the Issues Involved in Collecting and Interpreting Online and Offline Data? *Internet Inquiry: Conversations About Method*, pp.33–53.

Pearrow, M. (2000). *Web site usability handbook with CD-ROM*. Rockland, Mass.: Charles River Media, Inc.

Petrie, H., Hamilton, F., King, N., and Pavan, P. (2006). Remote usability evaluation with disabled people. *Proceedings of CHI 2006, ACM*, pp. 1133–1141.

Preece, J., Rogers, Y., Sharpe, H., Benyon, D., Holland, S., and Carey, T. (1994), *Human-Computer Interaction*, Addison-Wesley.

Punch, K.F. (2005). *Introduction to Social Research: Quantitative and Qualitative Approaches*. London: SAGE Publications Ltd.

Quesenbery, W. (2001, May). What does usability mean: Looking beyond "ease of use". *Annual conference-society for technical communication*, 48, pp. 432-436.

Rhemtulla, M., Brosseau-Liard, P.É. and Savalei, V., (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), p.354.

Rubin, J., and Chisnell, D. (2008). *Handbook of usability testing* (2nd ed.). Indianapolis: Wiley.

Rudd, J., Stern, K., and Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *Interactions*, 3(1), pp. 76–85.

Salvucci, D. D., Taatgen, N. A., and Borst, J. P. (2009, April). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1819–1828.

Sasangohar, F., Donmez, B., Trbovich, P., and Easty, A. C. (2012, September). Not all interruptions are created equal: positive interruptions in healthcare. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA: SAGE Publications, 56(1), pp. 824–828.

Sauer, J., and Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics*, 40(4), pp. 670–677.

Sauer, J., Seibel, K., and Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics*, 41, pp. 130–140.

Saunders, M., Lewis, P., and Thornhill, A. (2009). *Research methods for business students* (5th ed), London: Pitman.

Sauro, J. (2010). If you could only ask one question, use this one. Retrieved from <http://www.measuringu.com/blog/single-question.php> (last viewed April 20, 2016).

Sauro, J., and Dumas, J. S. (2009, April). Comparison of three one-question, post-task usability questionnaires. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 1599–1608.

Sauro, J., and Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research.* Elsevier / Morgan Kaufmann.

Scandura, A., and Williams, E. (2000). Research methodology in management: Current practices, trends and implication for future research. Academy *of Management Journal*, 143(6), pp. 1248–1264.

Schiffman, N., and Greist-Bousquet, S. (1992). The effect of task interruption and closure on perceived duration. *Bulletin of the Psychonomic Society*, 30(1), pp. 9–11.

Scholtz, J. (1999, January). A case study: developing a remote, rapid, and automated usability testing methodology for on-line books. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. (1999). HICSS-32. Abstracts and CD-ROM of Full Papers, IEEE,* pp. 11.

Scott, M. M. (2005). A powerful theory and a paradox: Ecological psychologists after Barker. *Environment and Behavior*, 37(3), pp. 295–329.

Sears, A., and Hess, D. J. (1999). Cognitive walkthroughs: Understanding the effect of task-description detail on evaluator performance. *International Journal of Human-Computer Interaction*, 11(3), pp. 185–200.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, USA: Houghton Mifflin Company.

Sheskin, D. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures* (5th ed.), Boca Raton, FL: Chapman and Hall / CRC.

Shneiderman, B., and Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. Boston, MA: Addison-Wesley.

Sierhuis, M., Clancey, W. J., and Van Hoof, R. J. (2007). Brahms: a multi-agent modelling environment for simulating work processes and practices. *International Journal of Simulation and Process Modelling*, 3(3), pp. 134–152.

Snowball, D. (1980). Some effects of accounting expertise and information load: An empirical study. *Accounting, Organizations and Society*, 5(3), pp. 323–338.

Snyder, C. (2003). *Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces*. San Francisco: Morgan Kaufmann.

Speier, C., Valacich, J.S., and Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2), pp. 337–360.

Speier, C., Vessey, I., and Valacich, J. S. (2003). The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decision Sciences*, 34(4), pp. 771–797.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive science*, 12(2), pp. 257–285.

Symonds, E. (2011). A practical application of SurveyMonkey as a remote usability-testing tool. *Library Hi Tech*, 29(3), pp. 436–445.

Tabachnick, B. G., and Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Boston: Pearson Education.

Tashakkori, A., and Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. *Applied Social Research Methods Series*, A Thousand Oaks, CA: Sage, 46.

Thompson, K. E., Rozanski, E. P., and Haake, A. R. (2004). Here, there, anywhere: remote usability testing that works. *Proceedings of the SIGITE '04. IT education—the state of the art.* Salt Lake City, USA, 28–30 October 2004. New York, USA: ACM, pp. 132–137.

Townsend, J. T., and Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological bulletin*, 96, pp. 341-401.

Trafton, J. G., Altmann, E. M., Brock, D. P., and Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5), pp. 583–603.

Tripathi, P., Pandey, M., and Bharti, D. (2010). Towards the identification of usability metrics for academic Websites. *The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, pp. 393–397.

Trivedi, M. C., and Khanum, M. A. (2012). Role of context in usability evaluations: A review. *Advanced Computing: An International Journal*, 3(2), pp. 69–78.

Tullis, T., and Albert, W. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. (2002). An empirical comparison of lab and remote usability testing of Web sites. *Proceedings of Usability Professionals Association Conference, Orlando, USA*, July 2002, viewed 30 March 2014, <http://home.comcast.net/~tomtullis/publications/RemoteVsLab.pdf>.

Tung, L.L., Xu, Y., and Tan, F.B. (2009). Attributes of Web Site Usability: A Study of Web Users with the Repertory Grid Technique. *International Journal of Electronic Commerce*, 13(4), pp. 97–126.

Tyler, S. W., Hertel, P. T., McCallum, M. C., and Ellis, H. C. (1979). Cognitive effect and memory. *Journal of Experimental Psychology: Human Learning and Memory*, 5, pp. 607–617.

Usability Tools (2016) Retrieved from usabilitytools.com in June 2016.

Van den Haak, M. J. and de Jong, M. D. T. (2005). Analyzing the interaction between facilitator and participants in two variants of the think-aloud method. *International Professional Communication Conference, 2005.*

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management*, 35(6), pp. 819–837.

Virzi, R.A., Sokolov, J.L., and Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. *Conference Proceedings on Human Factors in Computing Systems: CHI 96*, pp. 236–243.

Watson, J. M., and Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic bulletin & review*, 17(4), pp. 479–485.

West, R., and Lehman, K. (2006). Automated summative usability studies: an empirical evaluation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 631–639.

Westbrook, J. I., Woods, A., Rob, M. I., Dunsmuir, W. T., and Day, R. O. (2010). Association of interruptions with an increased risk and severity of medication administration errors. *Archives of Internal medicine*, 170(8), pp. 683–690.

Winckler, M. A., Freitas, C. M., and de Lima, J. V. (2000). Usability remote evaluation for www. *CHI EA 2000: Proceedings of the CHI 2000 Extended Abstracts on Human factors in Computing Systems*, The Hague, The Netherlands, 1-6 April 2000, New York, USA: ACM, pp. 131–132.

Wixon, D. (2003). Evaluating usability methods: why the current literature fails the practitioner. *Interactions*, 10(4), pp. 28–34.

Woolrych, A., Hornbæk, K., Frøkjær, E., and Cockton, G. (2011). Ingredients and Meals Rather Than Recipes: A Proposal for Research That Does Not Treat Usability Evaluation Methods as Indivisible Wholes. *International Journal of Human-Computer Interaction*, 27(10), pp. 940–970.

Zajonc, R.B., 1965. Social facilitation. *Science*, 149, pp. 269–274.

Zaphiris, P., and Kurniawan, S. (2006). *Human-Computer Interaction Research in Web Design and Evaluation*. Idea Group Publishing.

Zhang, Z., Basili, V. and Shneiderman, B. (1998) 'An Empirical Study of Perspective-Based Usability Inspection', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(19), pp. 1346–1350. doi: 10.1177/154193129804201904.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. Delft, Netherlands: Delft University Press.

Zviran, M., Glezer, C., and Avni, I. (2006). User satisfaction from commercial web sites: The effect of design and use. *Information & Management*. 43, pp. 157–178.

Appendix A.CH3: Methodology

- **Exploring data**

This procedure aims to describe the characteristics of the test data. Before conducting the statistical analysis (e.g. t-test), it is important to check that none of the assumptions made by the individual tests are violated. Testing of assumptions usually involves obtaining descriptive statistics on the variables, which include the mean, standard deviation, range of scores, skewness and kurtosis for continuous variables, and frequencies for categorical variables.

It is important to consider how to deal with missing values in the statistical analysis. Three options are available with each statistical analysis test in SPSS: (1) exclude cases listwise; (2) exclude cases pairwise; or (3) replace the missing value with mean. Listwise indicates that if any case contains any missing data, it (the case) will be excluded from the analysis, which results in limiting the sample size. Pairwise indicates that the case will be excluded only if it is missing the data required for the specific analysis, which leaves the case available for other analysis that does not require that data. The problem with the third option—replacing the missing value with mean—is that it can severely distort the statistical analysis results, especially if there are a lot of missing values. As a result, in the present study, missing data have been pair-wisely excluded in the analysis process when applying the statistical tests.

- **Assessing data normality distribution and variation**

Many statistical techniques (e.g., t-test, ANOVA, correlation, etc.) assume that the distribution of scores on the dependent variable is normal. Normal is used to describe a symmetrical, bell-shaped curve, which has the greatest frequency of scores in the middle with smaller frequencies toward the extreme. Normality can be assessed to some extent by obtaining skewness and kurtosis values, where the value of skewness indicates the symmetry of the distribution, and the value of kurtosis provides information about the 'peakedness' of the distribution. If the distribution is perfectly normal, the values of skewness and kurtosis will equal 0. Positive skewness indicates a clustering of the scores to the left at low values and vice versa. Positive kurtosis values indicate that the distribution is rather peaked (clustered in the centre) with long thin tails, while negative kurtosis indicates a relatively flat distribution (in many cases, in the extreme). As large as the sample could be, the skewness will not 'make a substantive difference in the analysis' (Tabachnick and Fidell, 2013, p. 80). Kurtosis can result in underestimating the variance, but this risk is reduced with a large sample, such as 200+ cases (Tabachnick and Fidell 2013, p. 80). This can be inspected from the data histogram (actual shape of distribution),

Normal Q-Q plot (the observed value for each score is plotted against the expected value from the normal distribution), Detrended Normal Q-Q plot (actual deviation of the score forms a straight line), Boxplot (the distribution of the scores for the two groups and very useful to detect outliers). However, the Kolmogorov-Smirnov test (K-S test) is used to assess the normality of distribution. If the K-S test statistic is significant at $p \leq 0.05$, then we can infer that the normality of the distribution of data is violated. Assessing Homogeneity of Variance indicates the assumption that the spread of outcome scores (scores of the dependent variables) is roughly equal at different scores on the independent variable. For correlational analysis, graphs might be useful, while with groups of data, Leven's test is used. Leven's test examines the null hypothesis that the variances in different groups are equal. If Leven's test is significant at $p \leq 0.05$, then we can conclude that the variances are significantly different and therefore the assumed homogeneity of variances has been violated.

- **Spotting and manipulating outliers**

Most of the statistical techniques are sensitive to outliers. Outliers can be inspected from histograms, where they lie on the tails of the distribution, sitting on their own out on the extremes. They could also be inspected from Boxplot provided by SPSS. In the SPSS Boxplots, points are considered as outliers (indicated as a little circle with a number attached, where the number is the corresponding case index) if they extend more than 1.5 box lengths from the edge of the box. Extreme points (indicated with an asterisk, *) extend more than three box lengths from the edge of the box. It is important to check that the outlier's score is genuine, not just an error; if it is not a typo and is a genuine score, then a decision should be made regarding what to do with the score. There are some possible techniques for removing all extreme values from the data (Tabachnick and Fidell, 2013). A similar technique called 'trimming the data' indicates the deletion of a certain number of scores from extremes based on two rules: (1) a percentage-based rule; and (2) a standard deviation-based rule. Percentage-based trimming is based on a percentage of data that is specified by either trimmed mean or M-estimator, which is determined empirically (Tabachnick and Filed, 2013). The advantage of trimmed means (and variance) is that they are accurate even if the distribution is not symmetrical because trimming the end of the distribution will remove outliers and skew that bias the mean. In contrast, standard deviation-based trimming will keep the mean and standard deviation influenced by outliers, so the criterion (standard deviation trimming) used to reduce the outliers' impact has already been biased by them (Tabachnick and Filed, 2013). The problem with trimming in SPSS is that there is no simple way to do it; although it will be calculated, the outliers and extreme data will not be excluded and should be done manually. Another technique involves changing the outlier

value to a less extreme value, thus allowing the corresponding case to be included in the analysis without allowing the score to distort the statistics (Tabachnick and Fidell, 2013). This is called 'Winsorizing'; however, it is dependent on whether the score that has been changed is unrepresentative of the sample as a whole, which might bias the statistical model, in which case it is considered to improve accuracy (Tabachnick and Filed, 2013).

However, the researcher decided to remove the outliers and extreme scores from the file. Doing that manually is an overwhelming task, as removing the cases and including them in other statistical analyses that do need that data may cause mistakes, loss, or overlooking returning removed cases. As a result, the researcher decided to use validation rules with selection data commands. Data validation allows for exploring the concepts of logical conditions or rules, which are very important for data manipulation. Validation in SPSS is a two-stage process: (1) create one or more logical rules that define valid data, and (2) apply the rules to the dataset. Therefore, single-variable rules will be created to check that the values in the corresponding variable lie within pre-defined ranges. The pre-defined range is the range that includes the outliers' values. If the rule is true, then the corresponding case will be invalid. Then, the selection commands will be based on selecting the valid data only. Therefore, the underside biased data and will be excluded from the data set when the validation rules and selection commands. Sometimes, it is a bit tricky to define a common range that includes the outliers. It might be simpler to exclude the range that contains the outliers and/or extreme values. In such a case, one possible way to do that is to use select cases command alone, where a logical rule can be defined to exclude cases that stratify that rule. Once the underlined analysis is completed, cases can be deselected again and re-included in other statistical analysis tests. The advantage of using validation over selection case command is that multiple rules can be defined on the data set, which eliminates the need to recreate the rule and select and deselect the cases each time. As a result, it has been decided to base removing the outliers from the analysis processes on using validation rules when possible, or if it is not the case, on the 'selection cases' command (alone) as an alternative technique.

- **Manipulating data**

Sometimes, it is necessary to add up the scores from the items that make up each scale to yield an overall score, such as rating Likert questions and multiple-choice questions. This involves two steps: (1) reverse any negatively worded items; then (2) add together scores from all the items that make up the subscale or scale. Questions may be designed differently from each other—some worded positively and others negatively—to avoid response bias. A positive

direction scale indicates that high scores indicate high optimism, while a negative direction scale indicates that high optimism is toward the lower value. Thus, if the scale is designed to be a positive direction scale, then scores assigned to positively worded questions will have different meanings than those assigned to negatively worded ones. The high optimism in the negatively worded questions has a negative meaning; therefore, the scales for the negatively worded questions need to be reversed. After reversing any negatively worded items in the scale, the next step is to calculate the total scores for each subject. However, SPSS provides the capability to encode variables (based on the given values given by the analyser) and to calculate the total scale scores. This procedure was used (in this research) when analysing the responses of rating questions of the online usability study that was deployed in the usability test.

SPSS enables the reduction or collapse of the number of categories of a categorical variable that might be desired in some instances. This also allows for collapsing of continuous variables (e.g., age) into categorical variables or ranges to analyse variance, which is useful for some analysis or with very skewed distributions. For example, the sample can be divided into equal groups according to the participants' scores on some variables. Visual binning is used to identify the suitable cut-off points to break the corresponding continuous variable into a new categorical variable that has only the specified values corresponding to a number of the underlined variable ranges chosen. However, one needs to be careful about converting continuous variables into dichotomous or categorical variables. One example is the practise of doing a "median split," which puts those with scores above and below the median into two categories, but other methods of artificial categorization can be just as problematic. Generally, a great deal of useful information is discarded, but other statistical issues arise. However, the practise of dichotomizing continuous variables is still quite prevalent. A paper by MacCullum et al. (2002) is a superb overview of the problems and potentially serious consequences of this practise. As a result, none of these procedures were utilised during the analysis process.

- **Checking scales' reilibility**

The reliability of a scale can vary depending on the sample. Therefore, it is necessary to check that each of the scales is reliable with a particular sample. If the scale contains some items that are negatively worded, these items need to be reversed before checking reliability. Sometimes scales contain several subscales that may or may not be combined to form a total scale score. If necessary, the reliability of each of the subscales and the total scale will need to be calculated. SPSS provides the capability to check the reliability of scales. Inter-Item Correlation Matrix values should all be positive, which would indicate that the scale's items are measuring the

same underlying characteristic; the presence of any negative values means that some items have not been correctly reverse-scored. This can also be inspected from the negative values of the Corrected-Item Total Correlation. Cronbach's alpha value should also be checked, where the values above 0.7 are considered acceptable and values above 0.8 are preferable. The Corrected Item-Total Correlation indicates the degree to which each item correlates with the total score, where low values (less than 0.3) indicate that the item is measuring something different from the overall scale. However, if Cronbach's Alpha value is too low (less than 0.7) and incorrectly scored items have been identified and resolved, it may be necessary to consider removing items with low total correlations. On any items of the scale, if alpha of Item Deleted value is higher than the final alpha value obtained, then these items may be removed from the scale. Reporting the mean inter-item correlation value with small scales (e.g. less than 10) is sometimes difficult to derive a decent Cronbach's Alpha value, allowing values of the mean inter-item in a specific range to suggest strong relationships among the items; nevertheless, that is not the case in many scales.

- **Selecting Statistical Analysis Tests**

In choosing the right statistic, several factors need to be considered. These factors differ whether we are using a questionnaire or experiment to collect data. In our research, the online study comprises both experiments that administered questions. When considering which questions to ask, we considered the type of the scale used (if they were scale-based questions), the nature of the data collected for each question (the score values of the variables corresponding that question) with the assumptions of the statistical techniques used to analyse the data collected for that question. Statistically, in our experiments, we were interested in the differences between groups (the samples in different environments) and the relationship between the data collected by those different groups. In terms of experimental research, factors like the nature of the dependent and independent variables should be considered (e.g., number of correct responses, ratings, length of time, categorical types) and then considering the level of measurement of dependent and independent variables. For continuous variables, information regarding their distribution (whether they are normally distributed or badly skewed), the range of the scores should be collected. For categorical variables, information regarding how many subjects (cases) fall into each category (whether the groups equal or very unbalanced) and whether some possible categories are empty should be considered. For the next step, a decision is made whether the statistical tests should be one of the parametric or nonparametric statistical test groups. Such decision should be taken after checking the distribution of data, and homogeneity of variances as described earlier (Section 4.7.2), if the data does not meet the assumptions of

the test we wish to use, then either choice can be made: manipulating the data which may make us unable to justify what we are doing (biased and distorted data) or using nonparametric tests which are not as powerful as the parametric tests but on the other hand, they are less sensitive to the outliers and skewness of the data. Parametric tests use raw or transformed data in the analysis of data, whereas nonparametric tests use the ranks of the data and do not attempt to estimate a population parameter from a sample statistic.

From another perspective, the choice of the statistical tests is typically based on more general or simpler classification of the level of measurements into “continuous” and “categorical”. These two general classes of measurement relate to two general classes of statistical tests—those based on normal theory and those based on binomial theory. Normal theory plays an important role in statistical tests with continuous dependent variables, such as t-tests, ANOVA, correlation, and regression, and binomial theory plays an important role in statistical tests with discrete dependent variables, such as chi-square and logistic regression. Classification of the independent and the dependent variable as continuous or discrete determines the type of statistical test that is likely to be appropriate in a given situation (Table 3.2).

However, there is a longstanding debate about how to classify measurements and whether levels of measurement can be a successful guide to choose data analysis type (Townsend and Ashby, 1984). In reality, several other factors must be considered in deciding on the most appropriate and statistically accurate analysis, including the distribution of the dependent variable, whether it is count data, and sample size, among others (Newsom, 2019). However, a problematic situation can occur when discrete numerical values like count variables are present—for example, in this research, the number of page views, number of usability problems identified, and the number of distractions events. Deciding whether to consider these values as categorical or continuous is a tricky task, nevertheless, because these count values indicate a magnitude that is explained by those numerical values. Such values of a variable indicate the scores (given to/by) each case, not the number of cases under a certain category. The numerical values assigned to the count variables have an order, equal intervals, and an absolute zero that is meaningful. For example, the number of usability problems encountered can be measured on a continuous level of measurement because a zero number of problems encountered means no presence of problems (Newsom, 2019).

Another issue is how to analyse the scores of Likert-type scales. Although these scales are technically ordinal, most researchers treat them as continuous variables and use normal theory statistics with them. When there are five or more categories, there is relatively little harm in

doing this (Rhemtulla et al., 2012). Most researchers probably also use these statistics when there are four ordinal categories, although this may be problematic at times. Additionally, once two or more Likert or ordinal items are combined, the number of possible values for the composite variable begins to increase beyond the 5 categories. Thus, it is usual practise to treat these composite scores as continuous variables (Newsom, 2019).

For ordinal analyses, ordinal scales with few categories (2, 3, or possibly 4) and nominal measures are often classified as categorical and are analysed using a binomial class of statistical tests, whereas ordinal scales with many categories (5 or more), interval, and ratio, are usually analysed with the normal theory class of statistical tests. On the other hand, the contrast between categorical and continuous variables is oversimplification. However, there is a big grey area when there are 3 or 4 ordinal categories. There is likely to be some statistical power advantage to using ordinal statistics over binomial statistics, and there is likely to be some accuracy gained in the statistical tests for using ordinal statistics over normal theory statistics when there are few categories or for certain other data conditions. Although the distinction is somewhat fuzzy, it is often a very useful distinction for choosing the preferred statistical test, especially at the beginning of the analysis (Newsom, 2019). Considering all the above factors, a decision-making framework was designed when choosing a statistical test in a specific situation during the analysis phases of this research design (Figure 3.6). In a situation analysing categorical dependent variables, the assumptions of a minimum of expected cell frequency are greater than '5' scores or at least 80% of the cells have expected frequencies of equal or greater than '5' scores. The reason behind this is that the problematic assumption is that with the chi-square test, the sampling distribution of the test statistic has an approximate chi-square distribution. Yet, with the larger sample, this issue seemed to be resolved. However, if the assumption is violated with small samples, then the cell indicated a group of scores that satisfy one option of both the independent and dependent variables together; the cell is one of the cells that comprise the contingency table. For example, if we have two variables, each of which has two options, then we have a 2*2 contingency table (4 cells). The significance test of chi-square distribution will be inaccurate. In the present study, Fisher's exact test will be used. The intuition behind Fisher's exact tests is its ability to calculate the significance of the test statistics can be calculated exactly, rather than relying on an approximation that approaching the exact value as the sample size grows to infinity like with many statistical tests. However, if the assumption of the Chi-square test is not violated, the Chi-square test will be used.

Appendix A.CH4: Exploratory Study

A.CH4.1 Standard Web Access

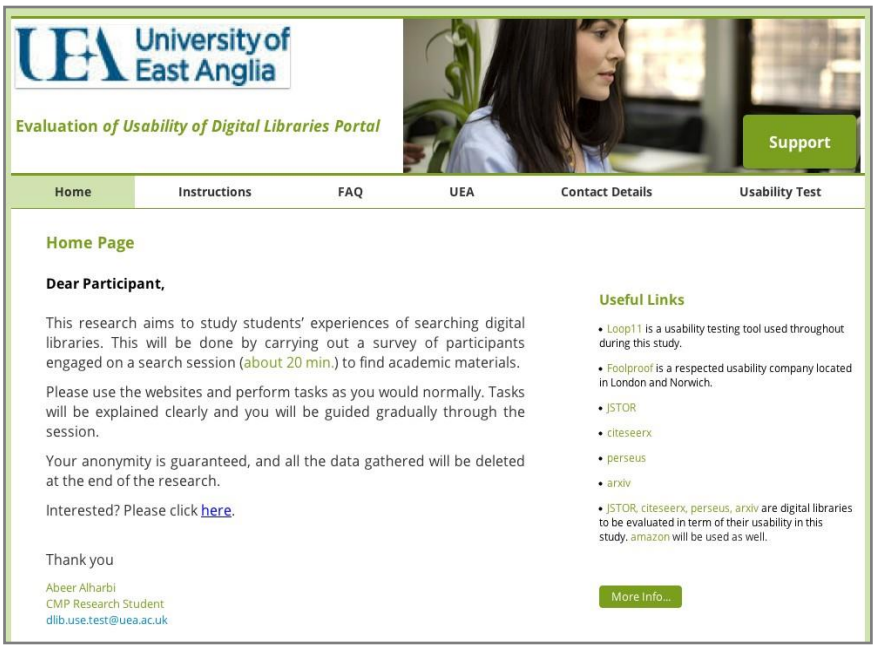


Figure A.Ch4.1: Standard web access page

A.CH4.2 Mobile Access

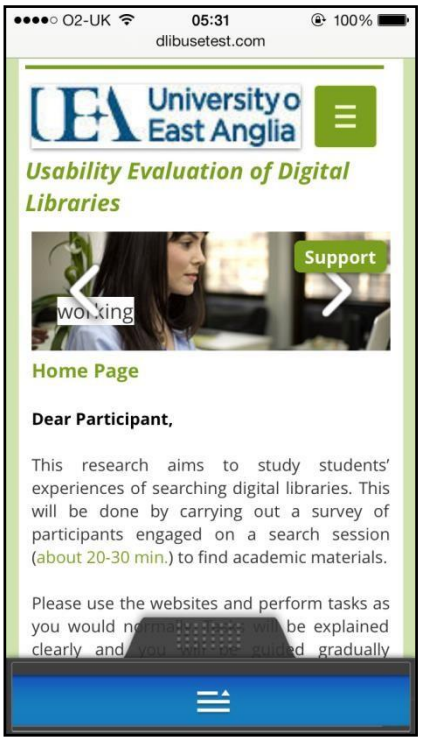


Figure A.Ch4.2: Mobile access layout

A.CH4.3 Loop11 Interface

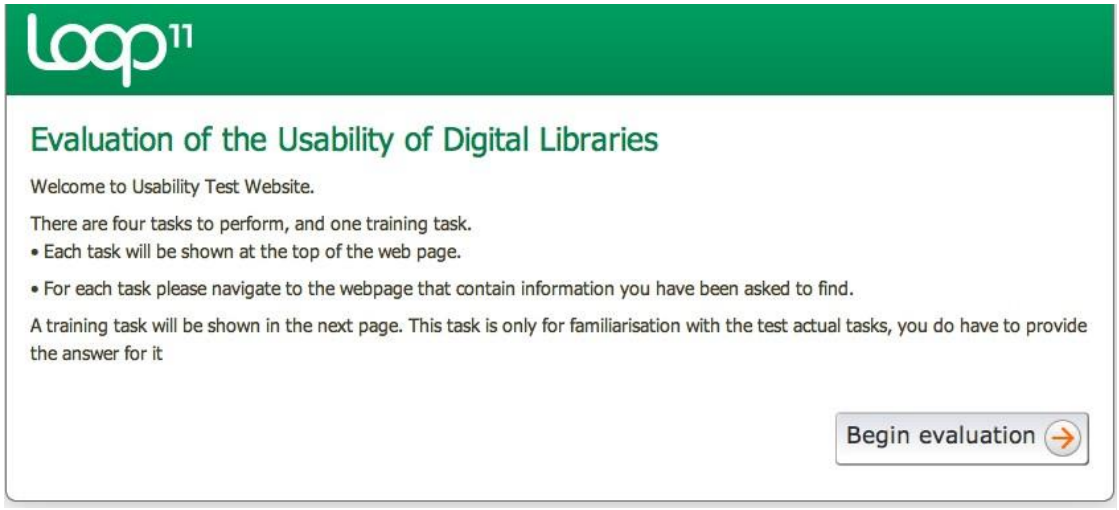


Figure A.Ch4.3: Loop11(welcome page)

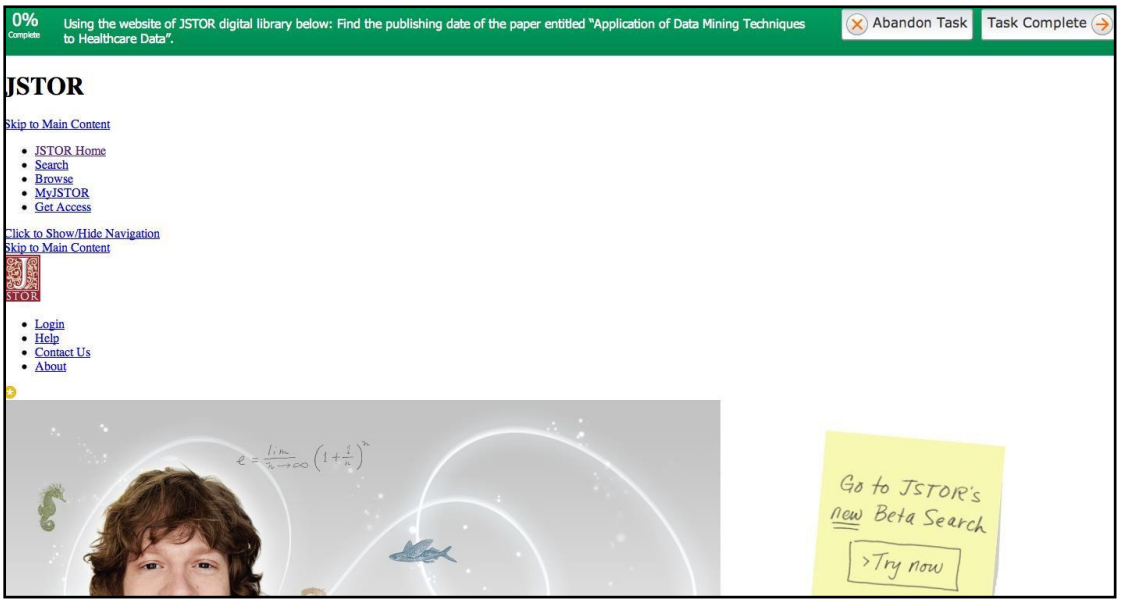


Figure A.Ch4.43: Loop11 (example of an experimental task)

Appendix A.CH5: Explanatory Study

A.CH5.1 Tasks Transcripts

Training Task

Training Task Object: Digital Public Library of America [DPLA]

URL: http://dp.la/

Training Task Description:

Visit the above listed digital library website and find the name of the publisher of the article entitled "Usability Testing for Voting Systems".

TASK A

Task A Object: The Universal Digital Library [UDL]

URL: http://www.ulib.org/

Task A Description:

Visit the above listed digital library website and find how many pages are in the English version of a book by "ALAN" about "Climate Change".

TASK B

Task B Object: Perseus Digital Library

URL: http://www.perseus.tufts.edu/hopper/

Task B Description:

Visit the above digital library website and determine how many lines there are in William Shakespeare's poem, "The Phoenix and the Turtle".

Do not forget to quote the number of lines you found.

TASK C

Task C Object: Cornell University Library [arXiv.org]

URL: http://arxiv.org/

Task C Description:

Visit the above listed digital library website and find how many figures are illustrated in the paper with the terms "AKARI" and "Luminosity" in its title. One of its authors is "Shuji Matsuura". The paper was published between 2010 and 2014 and it is 10 pages long.

Do not forget to quote the number of figures you have found.

TASK D

Task D Object: Amazon.co.uk

URL: https://www.amazon.co.uk/

Task D Description:

Visit the website Amazon.co.uk to find name of the publisher of the 3rd edition of the book entitled *Academic Writing for Graduate Students* written by Swales and Feak.

A.CH5.2 Test Advertisement

Usability Evaluation of Digital Libraries

* **The participation is available until the 4th of March ___**

Dear participant,

I am currently studying toward a PhD in Computing Sciences at UEA, and as part of my PhD thesis, I am doing a research project on website usability. I am looking for participants to take part in an experiment.

What can I expect if I participate?

You will be engaged to search digital libraries website(s) to find materials, express your experience with these websites, and rate their usability. Your performance data will be collected and then analysed in order to answer our research questions. There will also some questions that you will simultaneously fill in while performing the test search session. This brief questionnaire will enable you to describe your experience with the test and to provide some typical demographic information.

Important:

- You will not be asked your name, and all data will be kept confidential and anonymous.
- No risks are associated with the study.
- **A £7 Amazon e-mail voucher will be sent to you after completing the test (subject to availability).**
- You should perform as you normally would, and your performance will not affect the incentive given.

How long will it take?

A session should take approximately **15–30** minutes (depending on each participant).

Interested?

If you are interested, please send an-e-mail to dlib.use.testing@uea.ac.uk with the subject of ***"Interested to participate in your usability testing study."***

When and where?

Once we receive your e-mail request for participation, you will receive an e-mail from us that will include further information about the test and how can you perform it. Remember, you can withdraw at any time from the test even while performing the test.

If you need additional information, please contact me at Abeer.Alharbi@uea.ac.uk or my supervisor Dr Pam Mayhew at P.Mayhew@uea.ac.uk.

Your contribution is highly appreciated.

Abeer Alharbi

Figure A.Ch5.1.1 Test advertisement e-mail transcript

The participation is available until the 4th of March 2016

Dear participant,

I am currently studying toward a PhD in Computing Sciences at UEA, and as part of my PhD thesis, I am doing a research project on website usability. I am looking for participants to take part in an experiment.

You will be engaged to search digital libraries website(s) to find materials, express your experience with these websites, and rate their usability. Your performance data will be collected and then analysed in order to answer our research questions.

There will also some questions that you will simultaneously fill in while performing the test search session. This brief questionnaire will enable you to describe your experience with the test and to provide some typical demographic information.

If you need additional information, please contact me at Abeer.Alharbi@uea.ac.uk or my supervisor Dr Pam Mayhew at P.Mayhew@uea.ac.uk.

Your contribution is highly appreciated.
Abeer Alharbi

Important to know:

- You will not be asked your name, and all data will be kept confidential and anonymous.
- No risks are associated with the study.
- A £7 Amazon e-mail voucher will be sent to you after completing the test (subject to availability).

Interested?

Please send an-e-mail to: dlib.use.testing@uea.ac.uk with the subject of ***"Interested to participate in your usability testing study."***

When and where?

Once we receive your e-mail request for participation, you will receive an e-mail from us that will include further information about the test and how can you perform it.
Remember, you can withdraw at any time from the test even while performing the test.

Figure A.Ch5.2. Test advertisement flyer

A.CH5.3 Academic Specialisation

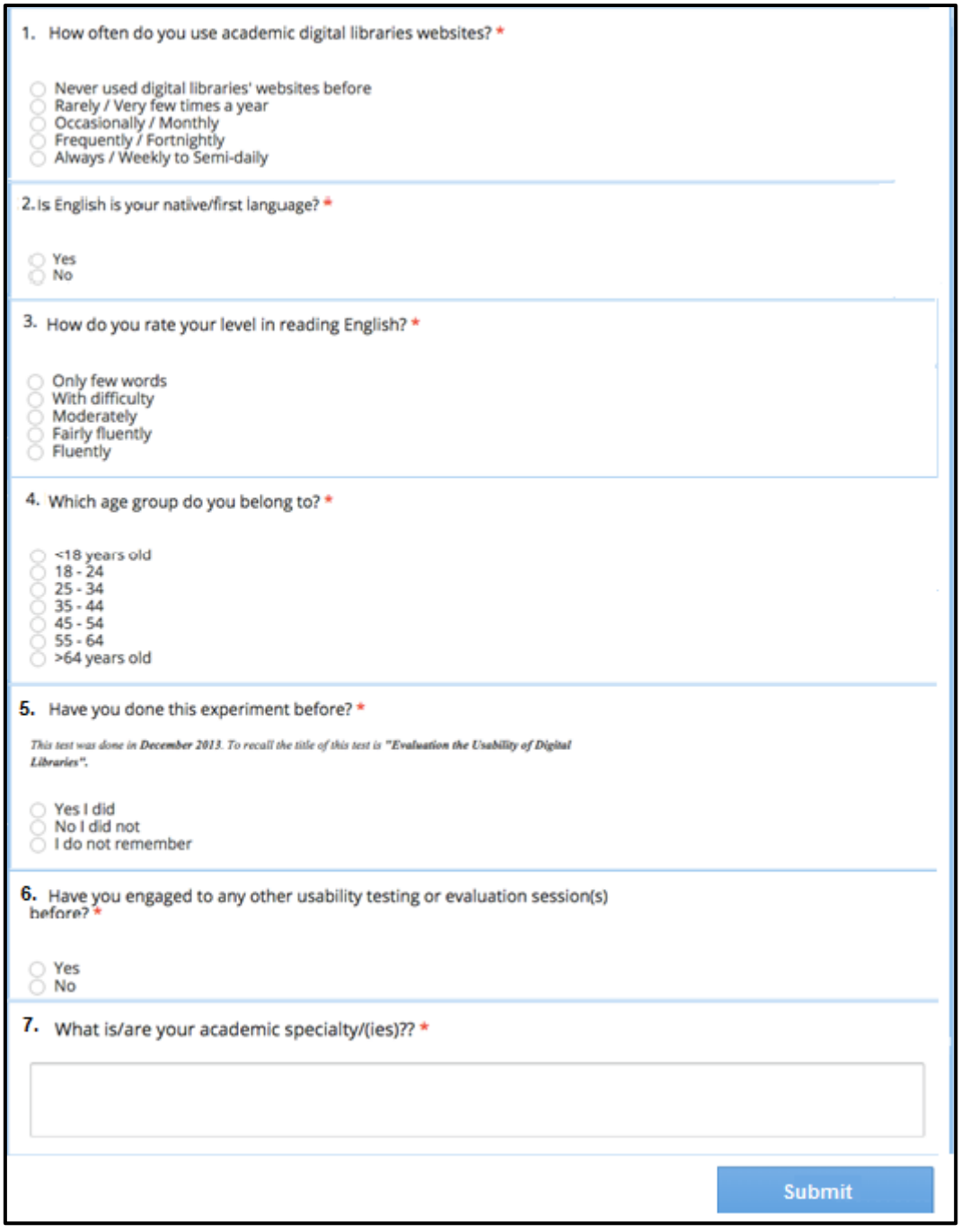
Table A.CH5.1. Academic Specialisations Classified as Text Oriented Subjects

Academic Speciality	Academic specialisations taught in UEA	
Text-oriented	• Adult Literacy, Lifelong learning and Development • Agricultural and ruler development • American studies • American history • American literature with creative writing • American and English literature • Applied Translations Studies • Archaeology, anthropology and art history • Biography and creative non-fiction • Broadcast Journalism: Theory and Practice • Climate change and International Development • Communications and Language studies • Conflict, governance and International Development Creative Entrepreneurship • Creative Writing • Culture, Literature and Politics • Development Economics • Drama • Early Modern History • Education • English Literature • English Literature and Drama • English Literature and Philosophy • English Literature with Creative Writing • English and American • Employment Law • Film studies • Film and English Studies • Film and History • Film and Television Studies • Film, Television and Creative Practice • Gender analysis and International Development • Geography • Geography and International Development • Globalisation Business and sustainable development • History • History of Art • History of Art and Literature • History of Art and Gallery and Museum Studies • History and History of Art • History and Politics • Information Technology and Intellectual Property Law • Intercultural communication with business management • Informational Commercial and Business Law • Informational Commercial and Competition • Law	• International Development • International Development with Anthology • International Development with Economics • International Development with Politics • International relations • International relations and politics • International Perspectives • International Public Policy and Public Management • International Public Policy, Regulation and Competition • International Relations • International Security • International Social Development • International Trade Law • Landscape history • Language and Intercultural Communication • Law • Law with American studies • Law with European Legal systems • Literary Translation • Literature and History • Media and Cultural Politics • Media Law, Policy and Practice • Medieval History • Modern and Contemporary Writing • Modern language(s) and management studies • Modern British History • Modern European History • Modern History • Modern languages • Mathematics Education • Media studies • Cultural Heritage and Museum Studies • Museum Studies • Philosophy • Philosophy and history • Philosophy and Literature • Philosophy and politics • Political, philosophy, language and communication studies • Politics • Politics and Media studies • Public Policy and Environment • Scriptwriting and Performance • Social Work • Society, culture and media • The Art of Africa, Oceania and the Americas • Theatre Directing: Text and Production • Translation and interpretation with modern languages • Translation and interpretation and modern language

Table A.CH5.2. Academic Specialisations Classified as Mathematically Oriented Subjects

Academic Speciality	Academic specialisations taught in UEA	
Mathematically oriented	• Accounting and Finance • Accounting and Management • Actuarial sciences • Adult Nursing • Advanced Organic Chemistry • Advanced practitioner: Emergency Case Practitioner • Advanced practitioner: Midwife • Applied computing Sciences • Applied Ecology and Conservation • Applied Ecology-International Program • Behavioural and Experimental Economics • Biochemistry • Biological Sciences • Biomedicine • Business Economics • Business Finance and Economics • Business Finance and Management • Business Information Systems • Business Management • Business Statistics • Brand leadership • Chemical Physics • Chemistry • Child nursing • Climate Change • Clinical Research • Clinical Research NIHR • Clinical Psychology • Coloproctology • Cognitive Neuroscience • Cognitive Psychology • Computer Graphics, Imaging and Multimedia • Computer Systems Engineering • Computer Science • Development Science • Ecology • Economics • Economics and Accountancy • Economics and International Finance and Trade • Economics and International Relations • Economy of Money, Banks, and Capital Markets • Energy Engineering • Energy Engineering with environmental Management • Engineering • Enterprise and Business Creation • Environmental Assessment and Management • Environmental Earth Sciences • Environmental Geography and Climate Change	• Environmental Geography and International Development • Environmental Geophysics • Environmental Sciences • Environmental science Finance and Economics • Finance and Management Forensic and Investigative Chemistry • Health Economics • Health Research • Human Resource Management • Industrial Economics • Information Systems • International Accounting and Financial Management • International Business Economics • International Business Finance and Economics • Investment and Financial Management • Knowledge Discovery and Data Mining • Leadership in Dementia Care • Leading Innovation for Clinical practitioner • Learning disabilities nursing • Management • Marketing • Marketing and Management • Mathematics • Mathematics with business • Media Economics • Medicine • Mental health nursing • Metrology and oceanography • Midwifery • Molecular Medicine • Molecular Biology and Genetics • Occupational Therapy • Oncoplastic Breast Surgery • Operations and Logistics Managements • Pharmacy • Paramedic Science • Pharmacology and Drugs Discovery • Physician Associate studies • Philosophy, Politics and Economics • Physiotherapy • Politics and Economics • Plants Genetics and Crop Improvement • Psychology • Quantitative Financial Economics • Regional Anaesthesia • Social Psychology • Speech and Language Therapy

A.CH5.4 Screening Questionnaire



1. How often do you use academic digital libraries websites? *

- Never used digital libraries' websites before
- Rarely / Very few times a year
- Occasionally / Monthly
- Frequently / Fortnightly
- Always / Weekly to Semi-daily

2. Is English is your native/first language? *

- Yes
- No

3. How do you rate your level in reading English? *

- Only few words
- With difficulty
- Moderately
- Fairly fluently
- Fluently

4. Which age group do you belong to? *

- <18 years old
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- >64 years old

5. Have you done this experiment before? *

*This test was done in **December 2013**. To recall the title of this test is **"Evaluation the Usability of Digital Libraries"**.*

- Yes I did
- No I did not
- I do not remember

6. Have you engaged to any other usability testing or evaluation session(s) before? *

- Yes
- No

7. What is/are your academic specialty/(ies)?? *

Submit

Figure A.Ch5.3. Participants' screening questionnaire using *UsabilityTools*

A. CH5.5 Online Portal

A. CH5.5.1 Desktop View

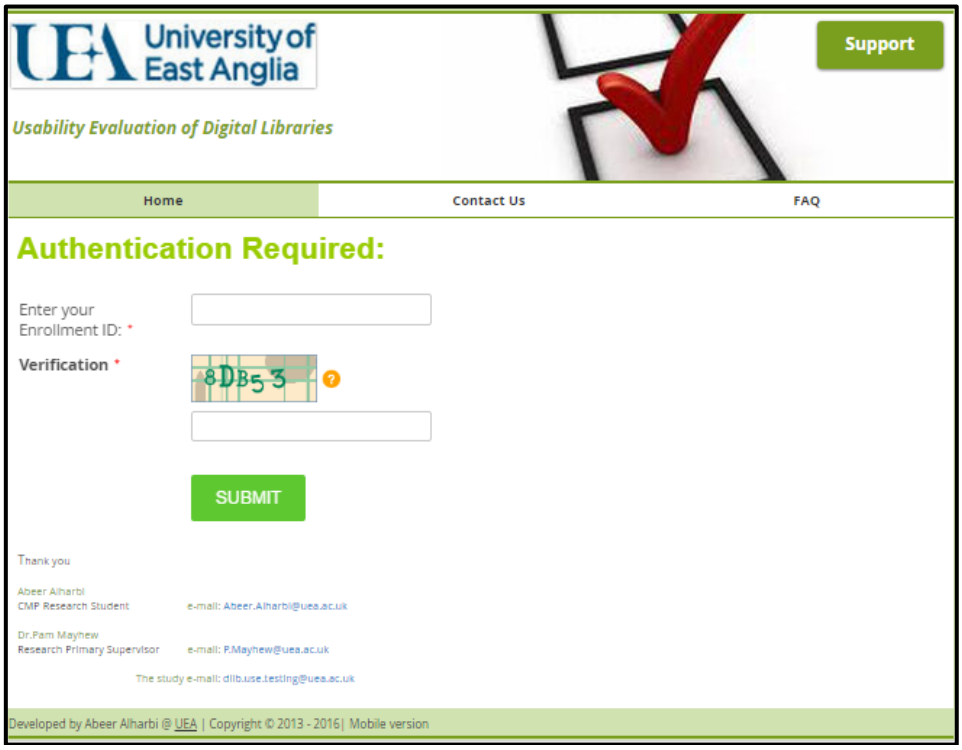


Figure A.Ch5.4. Online Portal – (home page, authenticating participant using Enrolment ID)

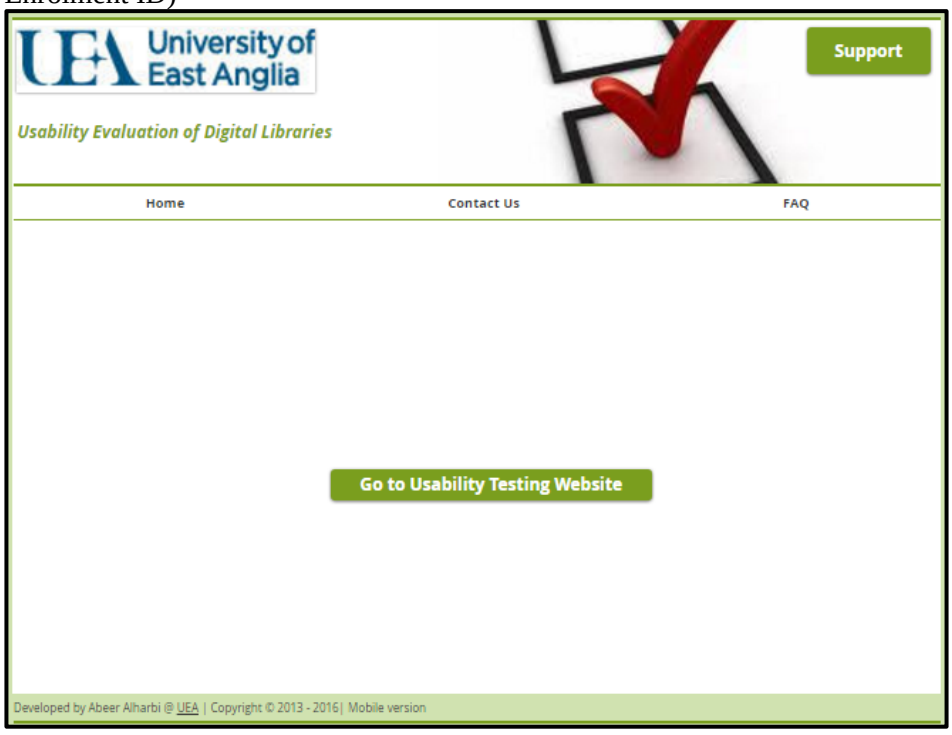


Figure A.Ch5.5. Online Portal – a unifying access page to the usability testing website, if the authentication was successful

A. CH5.5.2 Mobile View

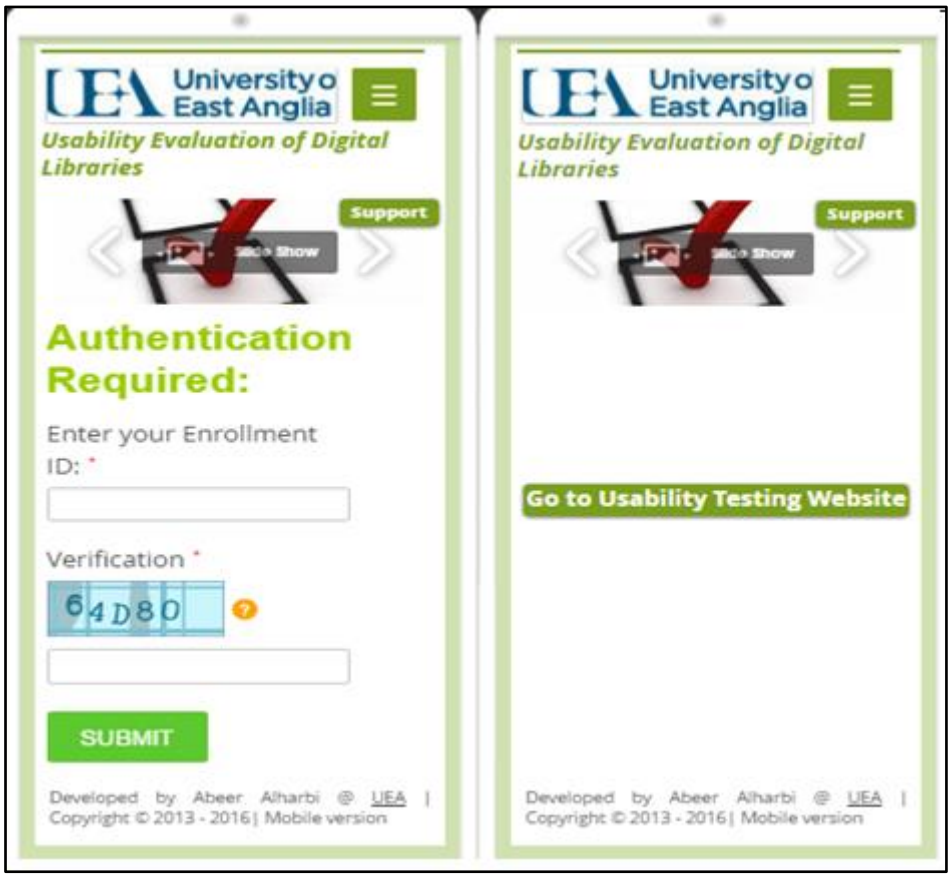


Figure A.Ch5.6. Online Portal – authentication, and unifying access page

A.CH5.6 UsabilityTools

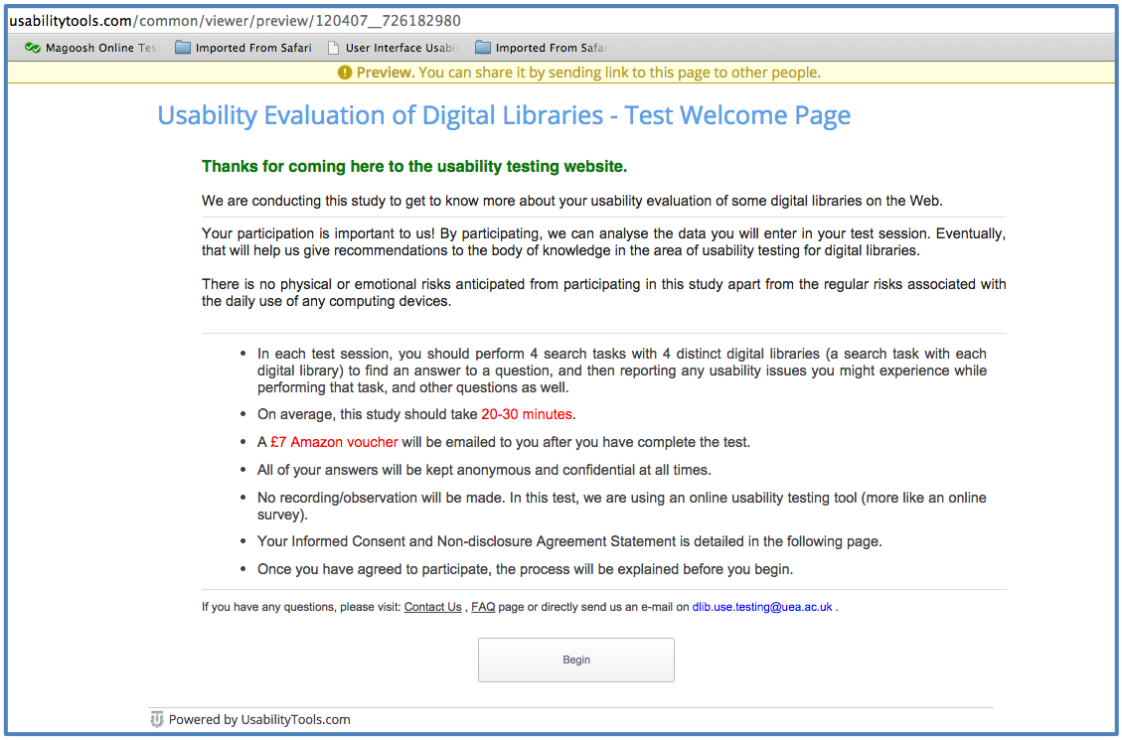


Figure A.Ch5.7. Online usability study (welcome page)

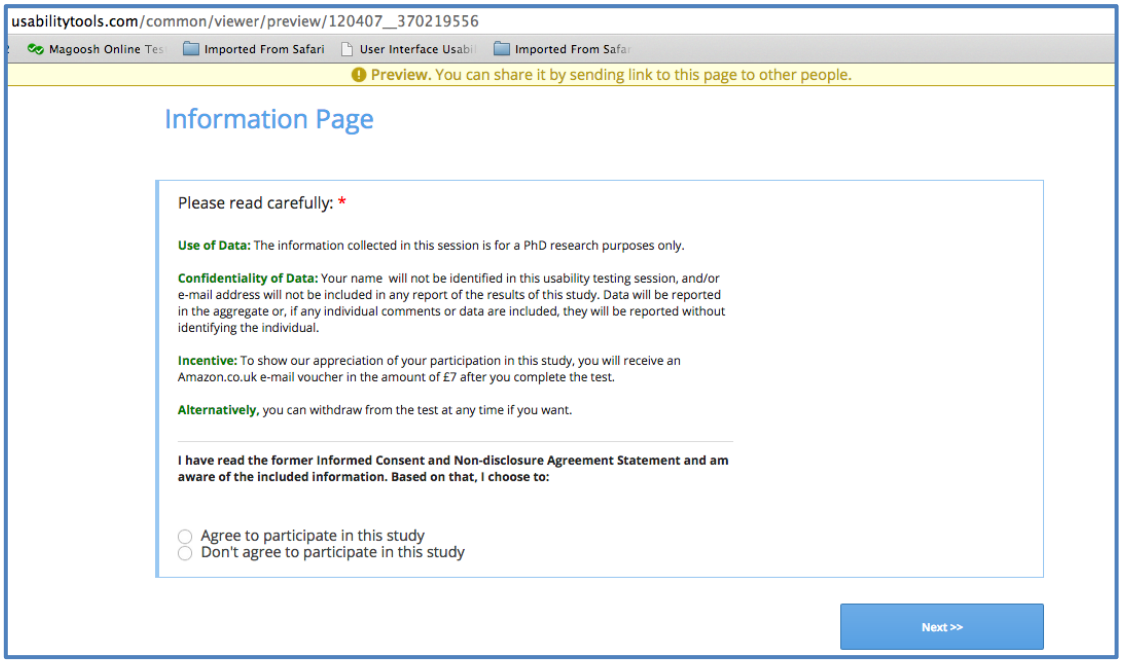


Figure A.Ch5.8. Online usability study (statement of informed consent)

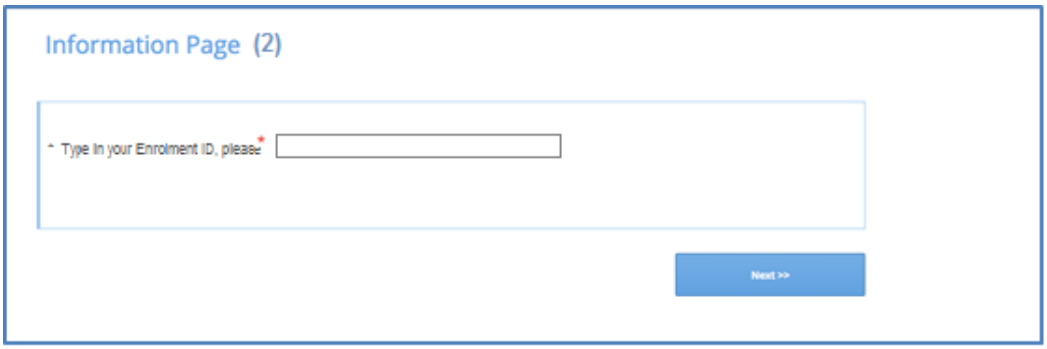


Figure A.Ch5.9. Online usability study (enrolment ID)

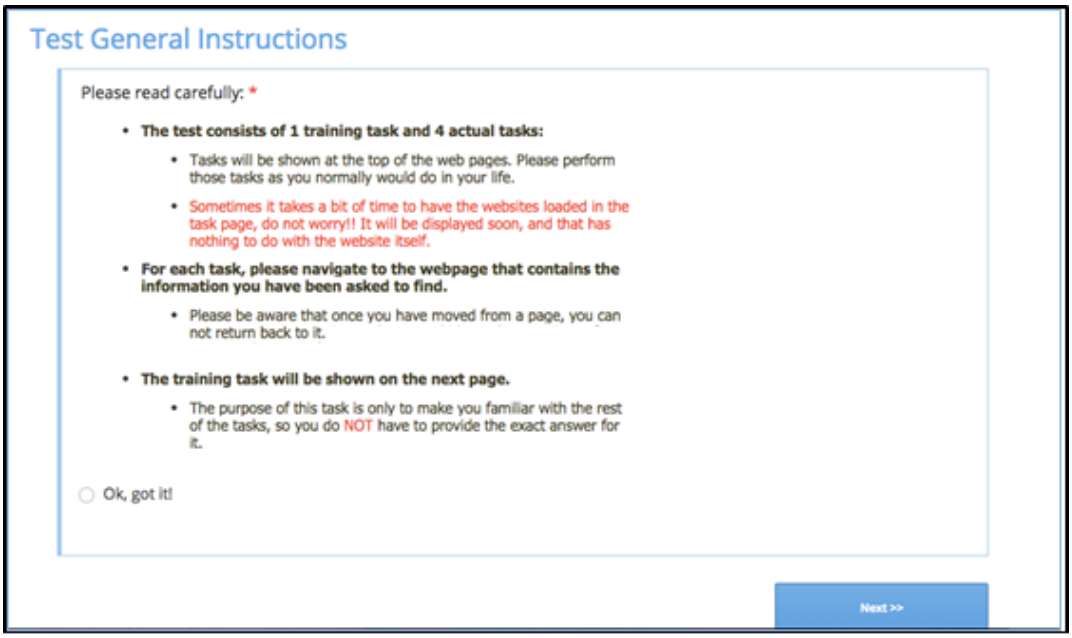


Figure A.Ch5.10. Online usability study (general instructions)

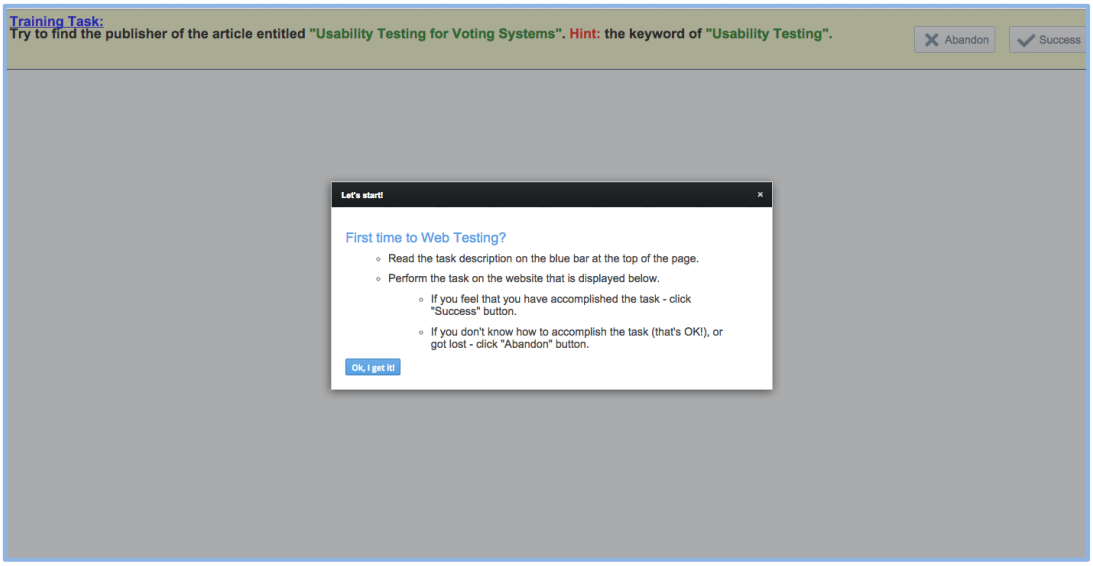


Figure A.Ch5.11. Online usability study (example of *UsabilityTools*-generated task instructions for the training task)

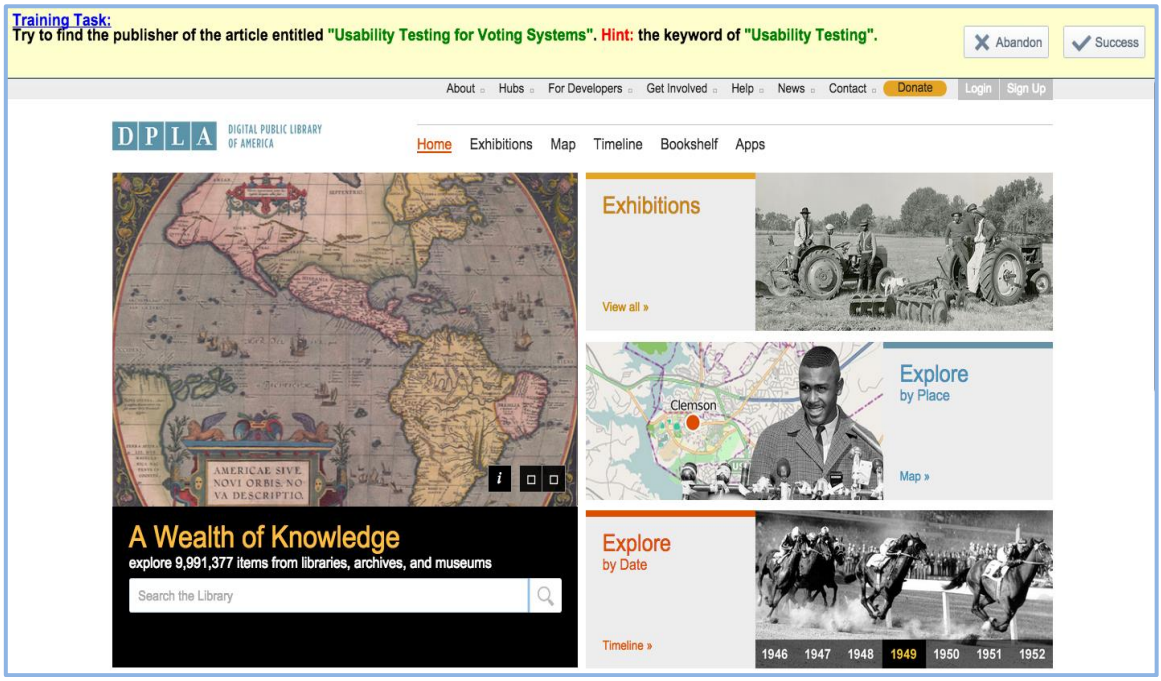


Figure A.Ch5.12: Online usability study (example of the presentation of tasks in *UsbailityTools*, Example given for Training Task)

Figure A.Ch5.13. Online usability study (self-assessment of success, example given for training task)

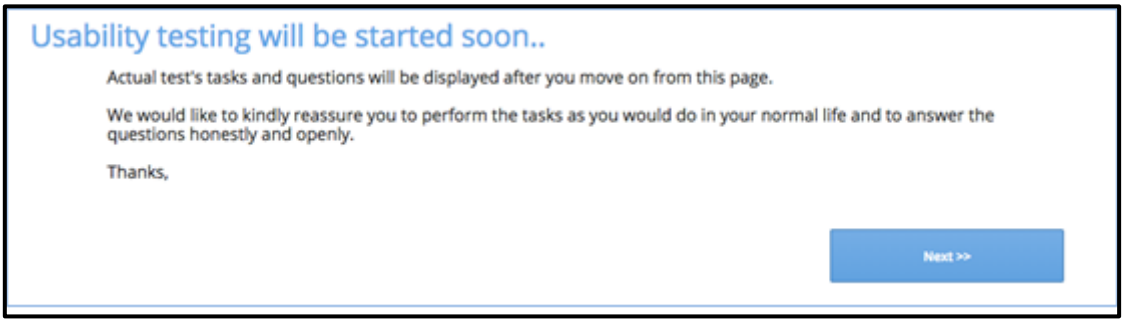


Figure A.Ch5.14. Online usability study (prompting participants before performing actual tasks)

Figure A.Ch5.15. Online usability study (example of an actual task)

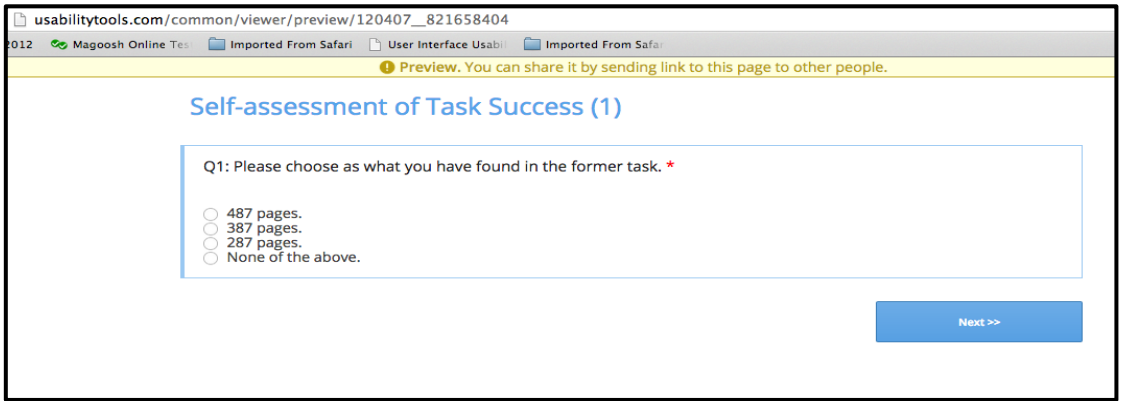


Figure A.Ch5.16. Online usability study (Example of Self-assessment of Task Success)

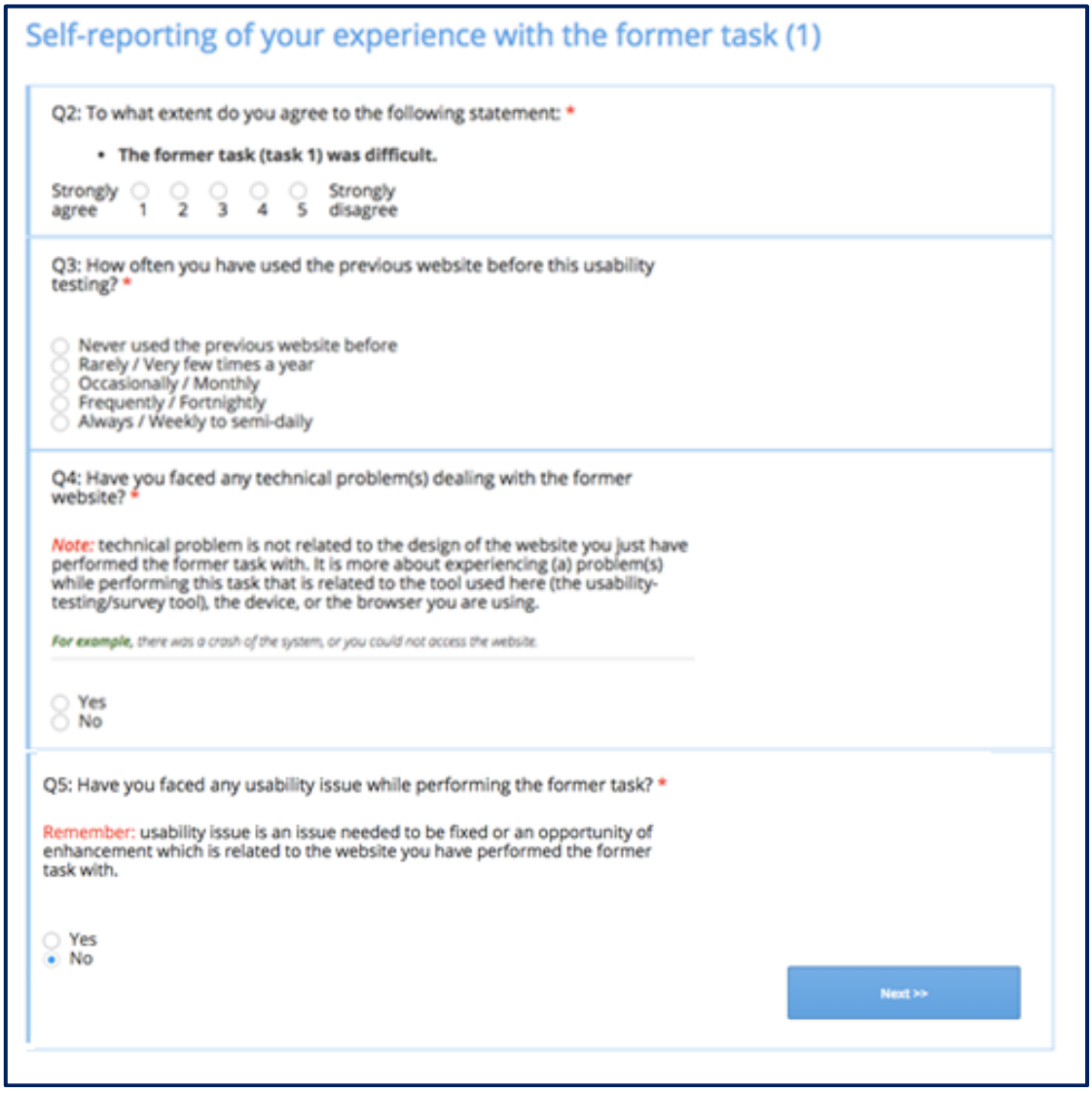


Figure A.Ch5.17. Online usability study (Self-reporting of participant experience with a task))

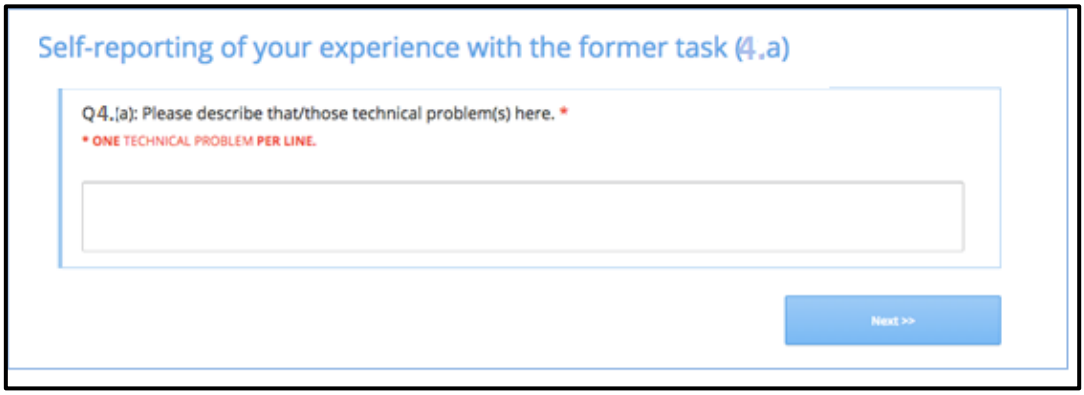


Figure A.Ch5.18. Online usability study (Example of Self-reporting of participant's experience with a task, branching question)

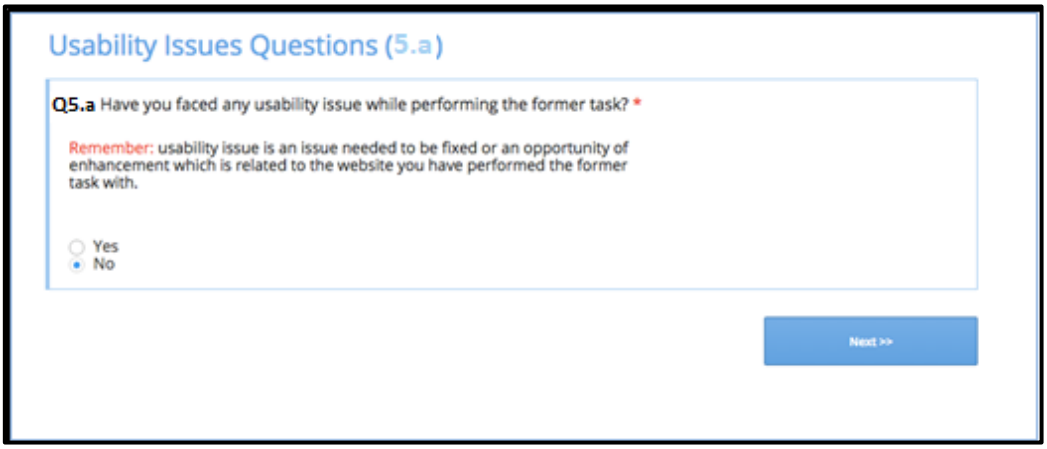


Figure A.Ch5.19: Online usability study (Example of a usability testing question (1) for a task)

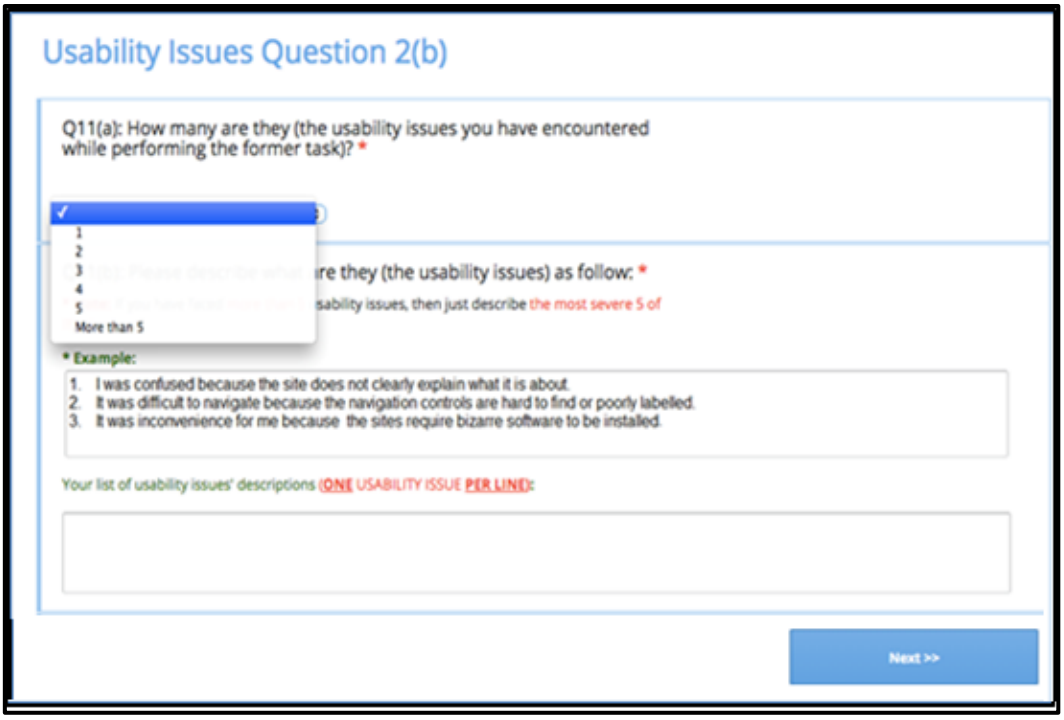


Figure A.Ch5.20: Online usability study (Example of usability issue question (2) for a task, "branching question")

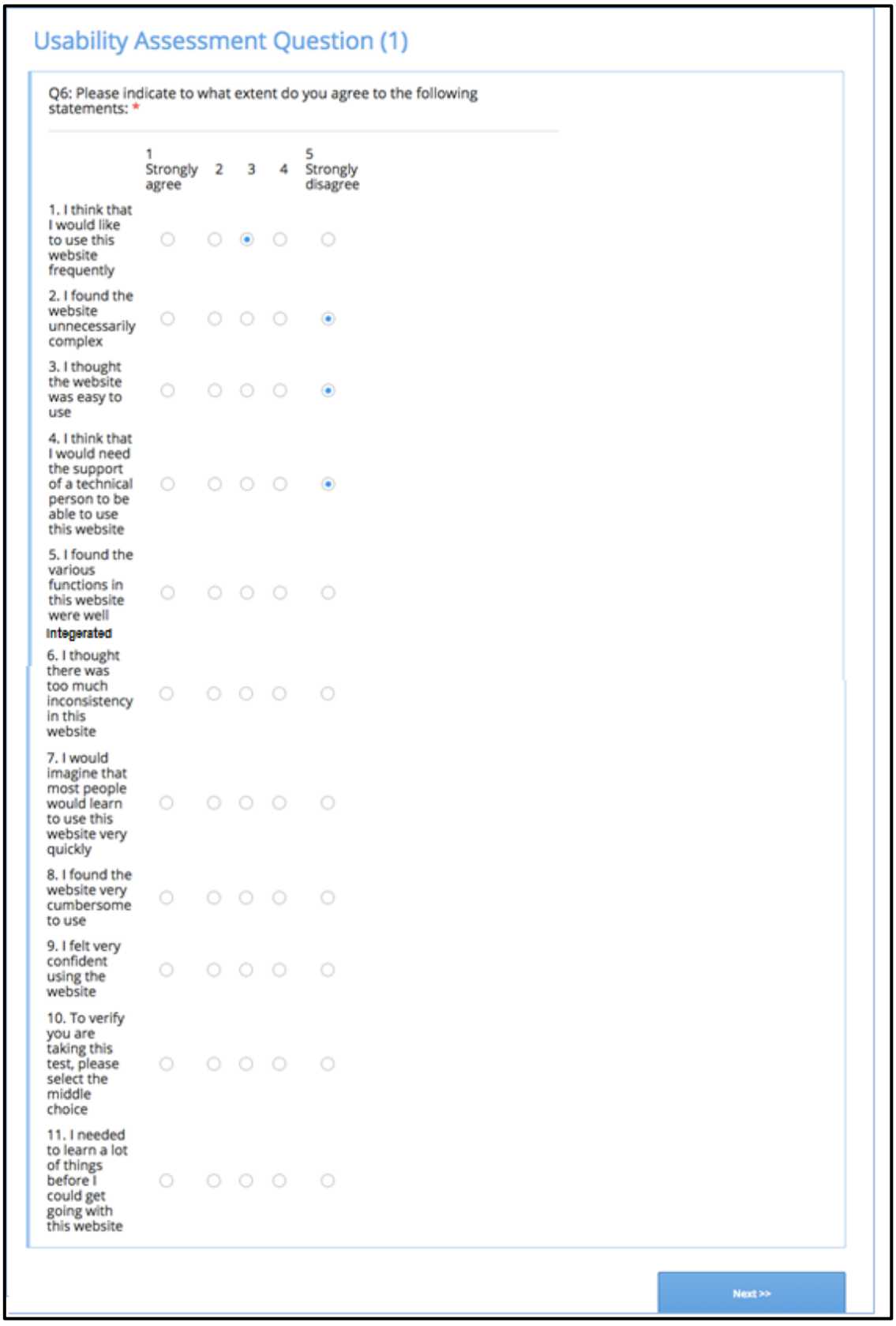


Figure A.Ch5.21. Online usability study (Example of SUS usability assessment standard questionnaire for a task)

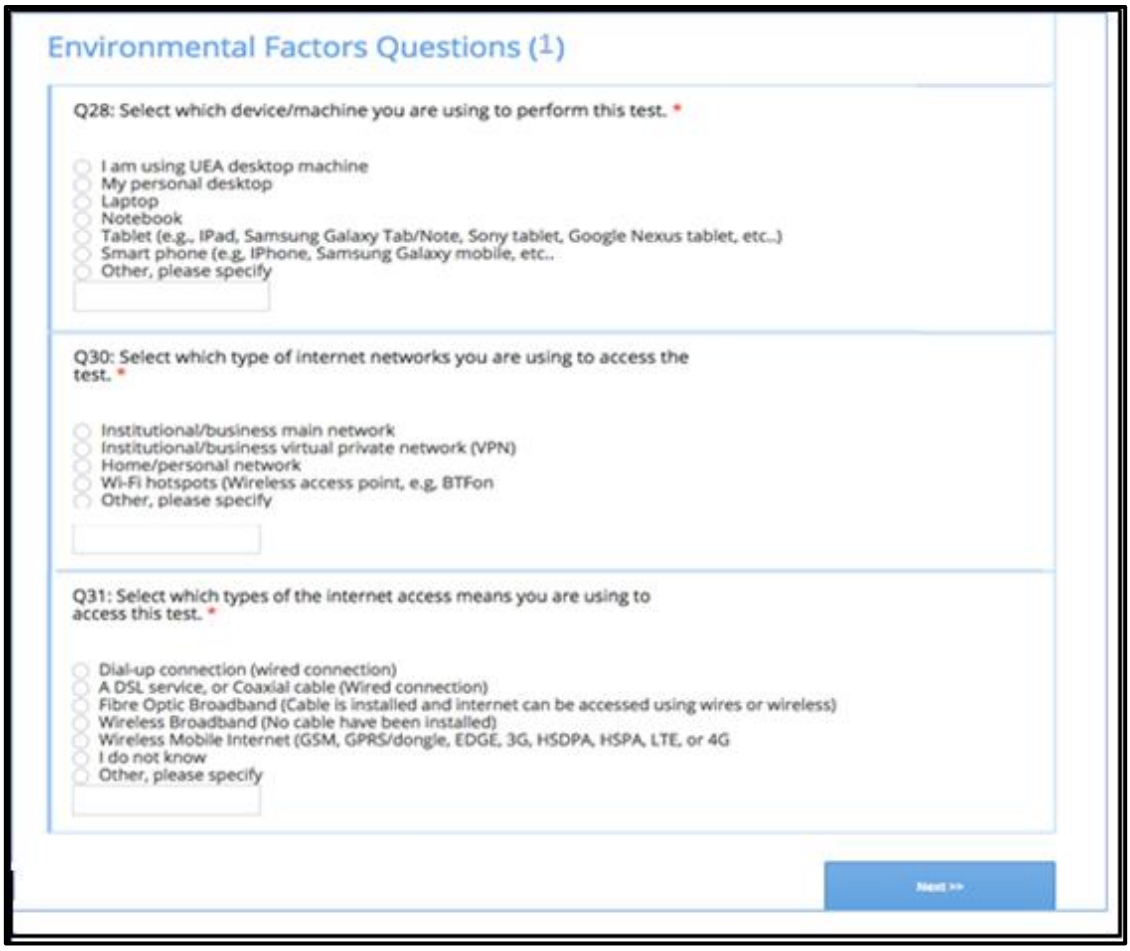


Figure A.Ch5.22. Online usability study (questions regarding contextual factors (1))

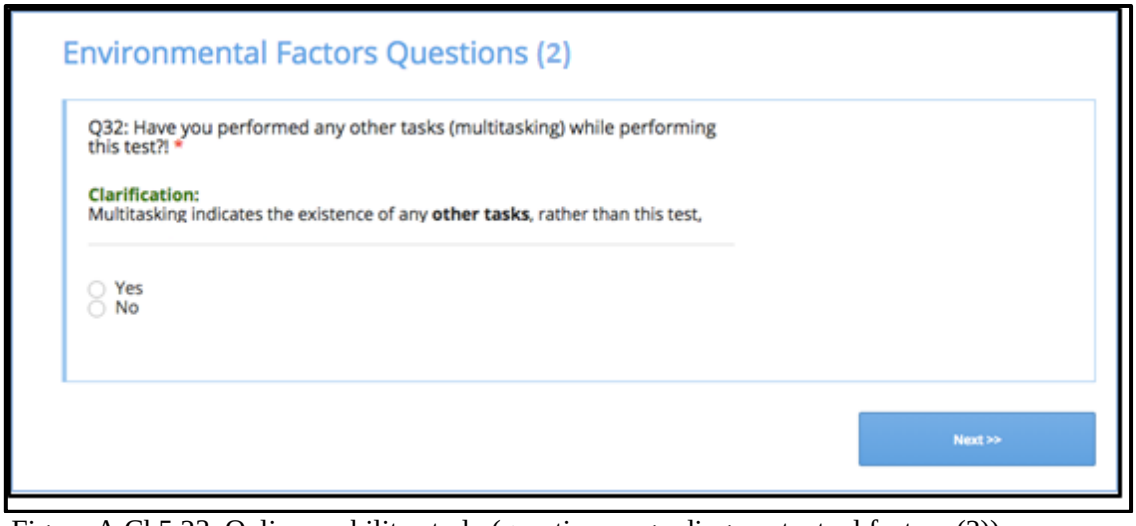


Figure A.Ch5.23. Online usability study (questions regarding contextual factors (2))

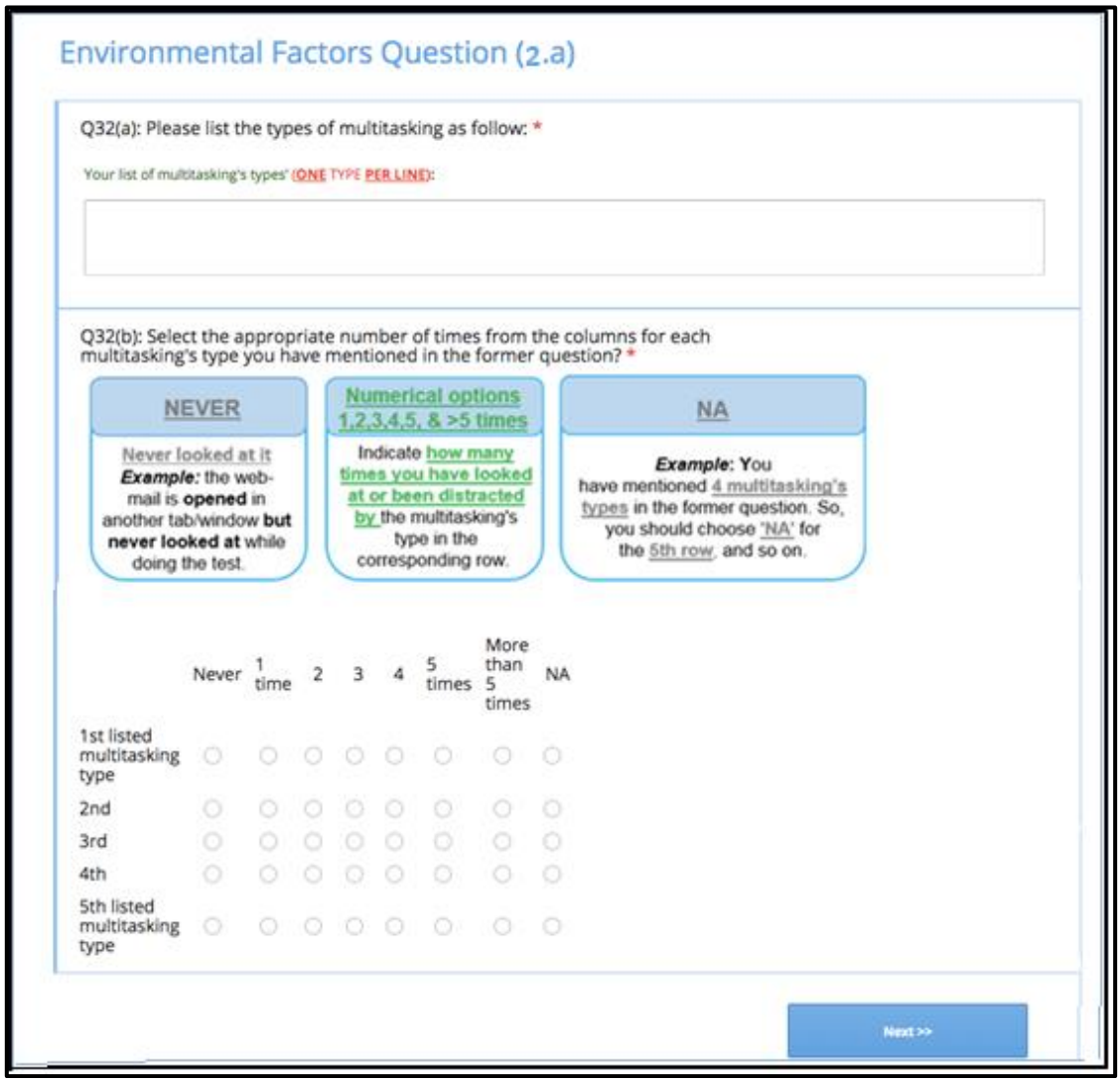


Figure A.Ch5.24. Online usability study (questions regarding contextual factors) (2.a), "branching question")

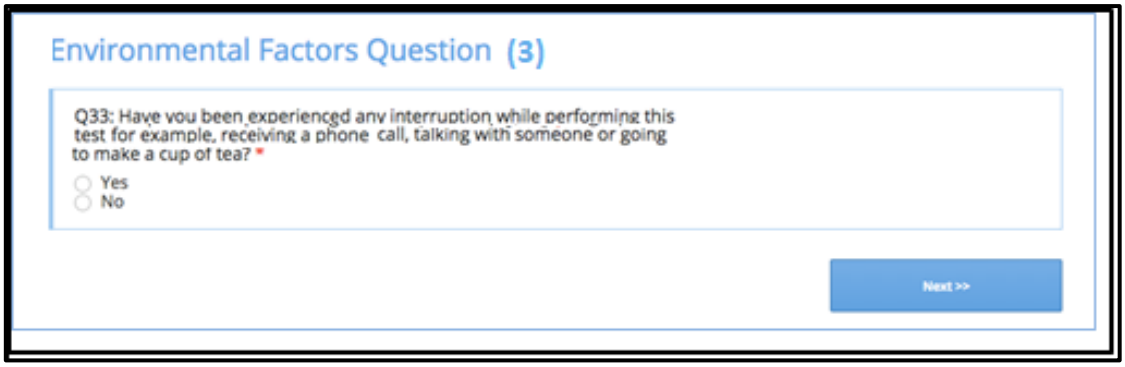


Figure A.Ch5.25. Online usability study (questions regarding contextual factors (3))

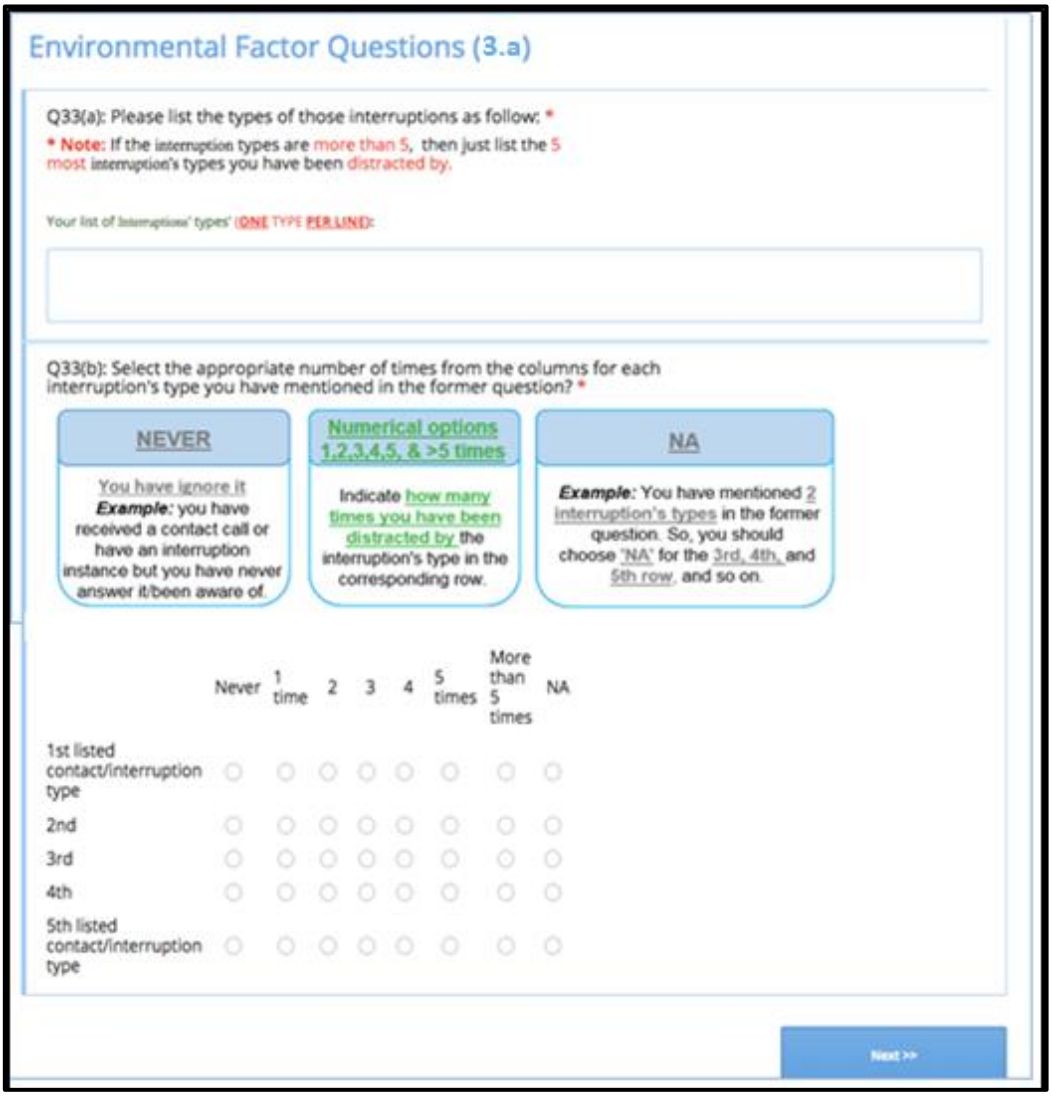


Figure A.Ch5.26. Online usability study (questions regarding contextual factors (3.a))

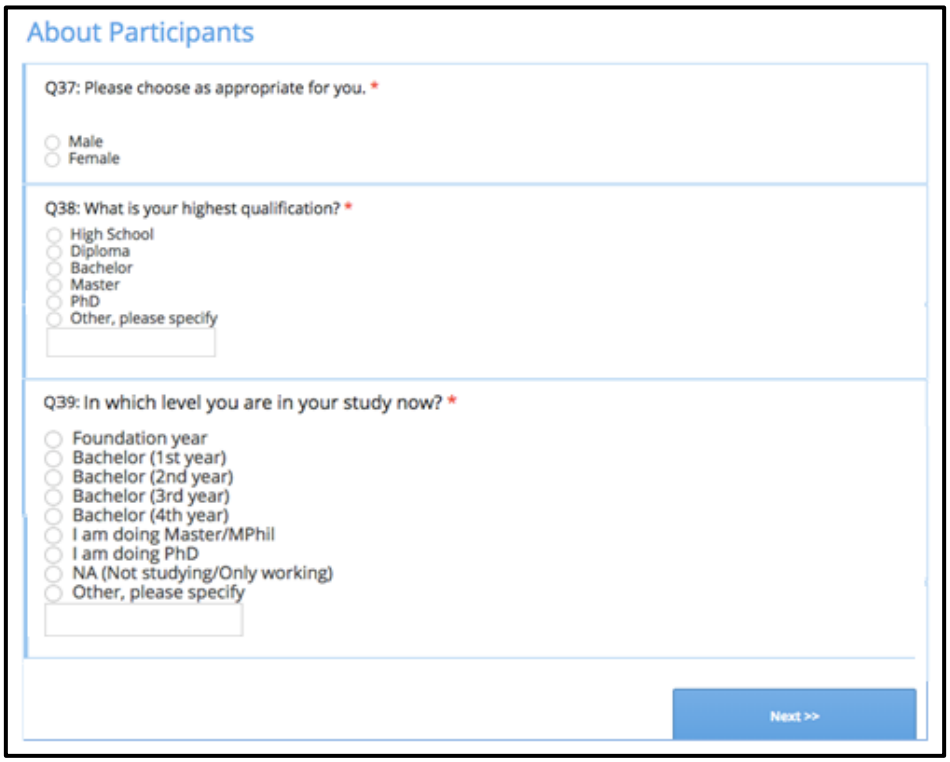


Figure A.Ch5.27. Online usability study (questions regarding participants' demographic)

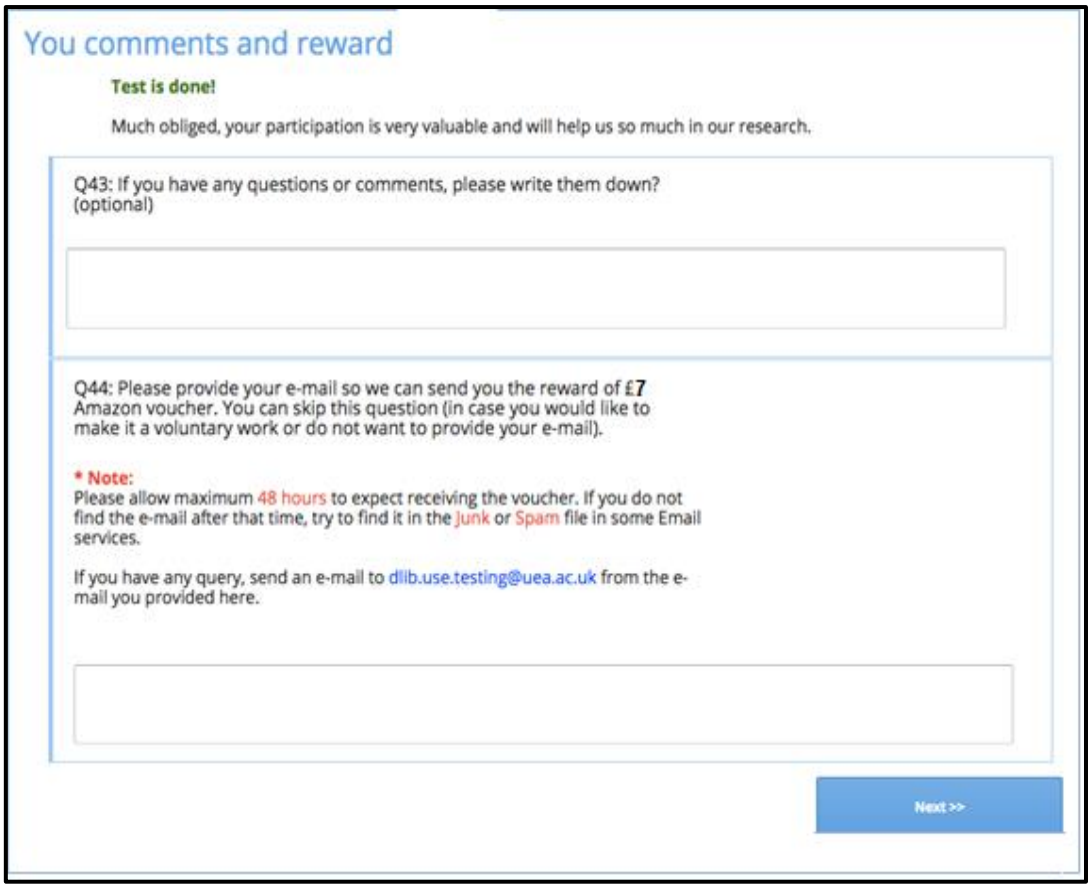


Figure A.Ch5.28. Online usability study (Comments and incentive)

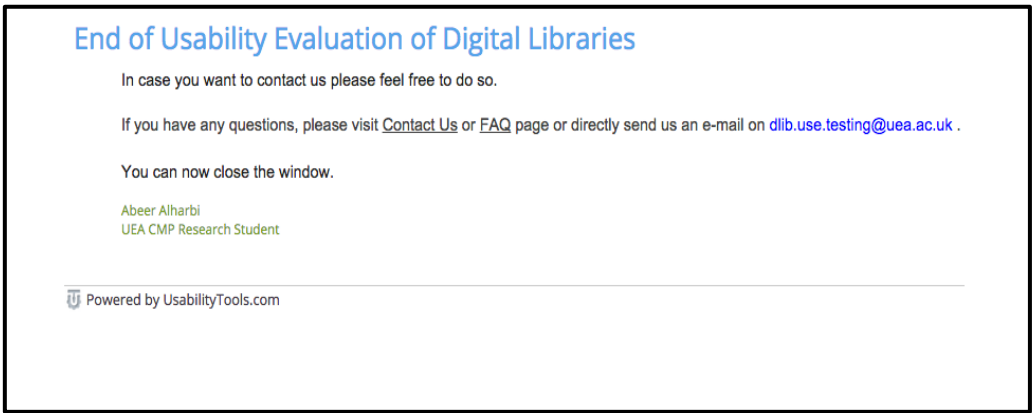


Figure A.Ch5.29. Online usability study (End page)

Appendix A.CH6: Validation Study

A.CH6.1 Information Sheet

Information Sheet

- Thank you for coming today. As may you already know my name is Abeer.
- The tasks mainly ask you to find items or information using the digital library website using the key information given. The key information is typically placed between single quotation marks (‘ ’) and the task should show you whether this information should be considered as subject, title, author, material type, keywords or catalogue section for example. This should help you know about what you are searching for—the key information, and using which search features—e.g., subject or author search or both.
- If the task is supported by a hint, please consider it while performing the task.
- You may use any feature of the website to find the answer for the task, but you may **NOT** use other websites (like Google or Yahoo), please stick to the website given.
- Please respond to the tasks naturally, as if you were using your preferred website.
- The task will be located on the side of the website page. You will read the task and try your best to perform it. You may hide it, restore it, or move it anywhere in the window. You may also copy and paste from it. Beside moving and hiding, the tasks have mini ‘exit’ button which you may use in case you decided to withdraw from the evaluation session.
- The task has two main buttons as well, “Task Complete” or “Abandon Task”. You should press one of them once you have finished the task.
- Please do **NOT** open other applications or programs while you are carrying out the tasks. Just stick to the browser window already open for you to perform the test.
- If the researcher needs to contact you for any reason, please **DO** attend her contact as soon as it happens using the smartphone provided on the desk, the researcher may call you or send you a message on the WhatsApp app installed in the smartphone provided.

Figure A.CH6.1: Transcript of the information sheet

A.CH6.2 Informed Consent Form

Informed Consent Form

The aim of this study is to evaluate a university library website. During the study, it will be necessary for me to record a number of things using screen capture software. Video and audio recordings will be valuable for the analysis if you agree for them to be taken. However, this recorded data will be stored securely on a password-protected computer in accordance with the UEA's data protection policy. The results of the analysis of this evaluation may be published, but all the data recorded will be anonymous. You can withdraw from this study at any time, in which case, recordings and notes taken will be destroyed.

Please tick the box below if you agree with, and sign below if you are happy to give your consent for the study to go ahead.

- I agree that my face and voice will be recorded. ☐ Yes ☐ No

Your signature:

! Signing this form indicates that you agree that your screen can be captured. ☐

Participant Name **Signature** **Date**

_____ _____ ___/___/2018

*If you would like to read to any reports or publications that result from this study, please tick the box. ☐

Contact Information:

	Researcher: Mrs. Abeer Alharbi	**Supervisor:** Dr. Pam Mayhew
Email address:	Abeer.Alharbi@uea.ac.uk	P.Mayhew@uea.ac.uk
Contact number:	07824016873	01603593334

Date: _____/_____/2018

Figure A.CH6.2: Transcript of the informed consent form

A.CH6.3 Screening Questionnaire

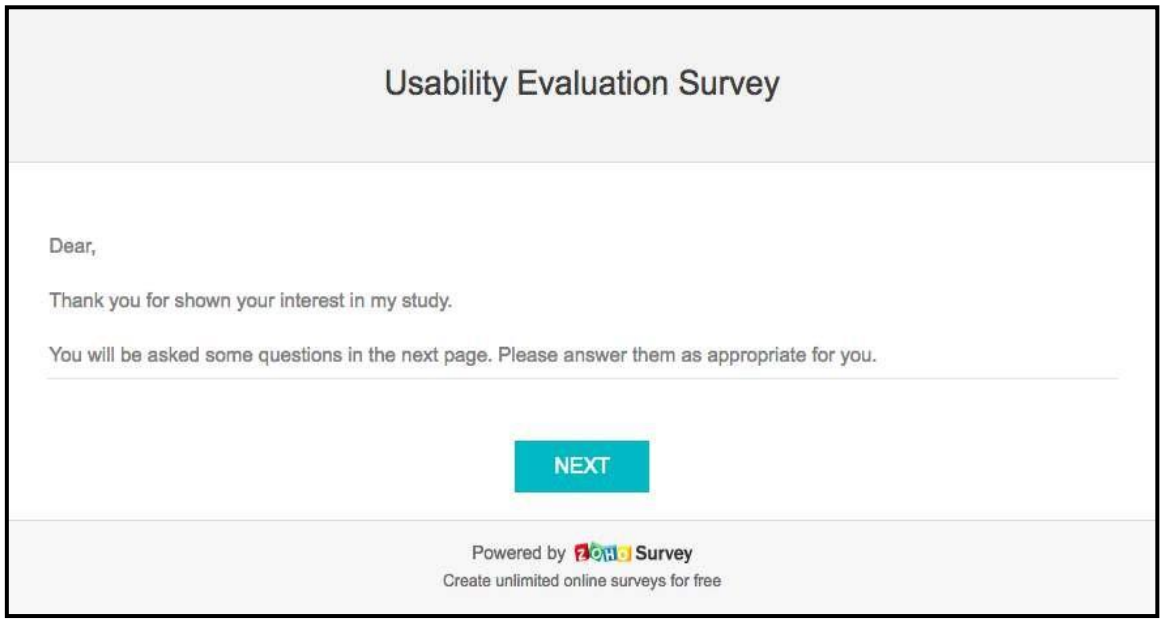


Figure A.CH6.3: Screening questionnaire (a)

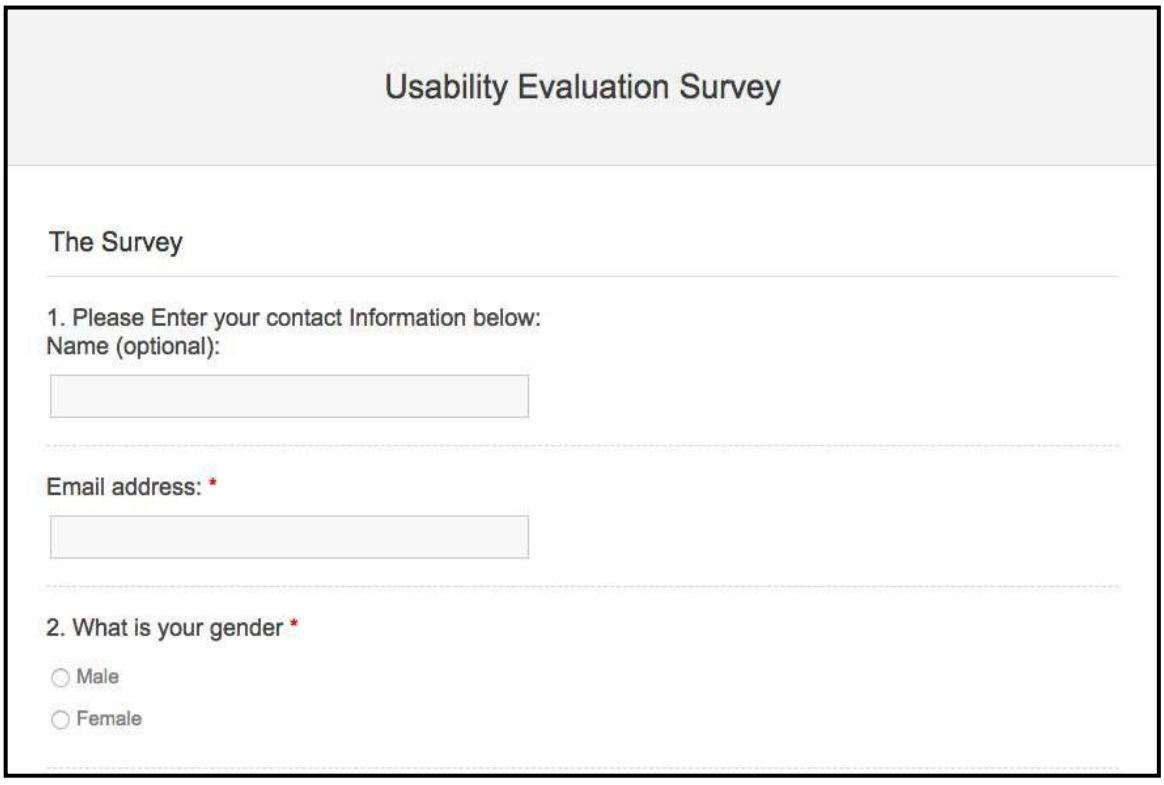


Figure A.CH6.4: Screening questionnaire (b)

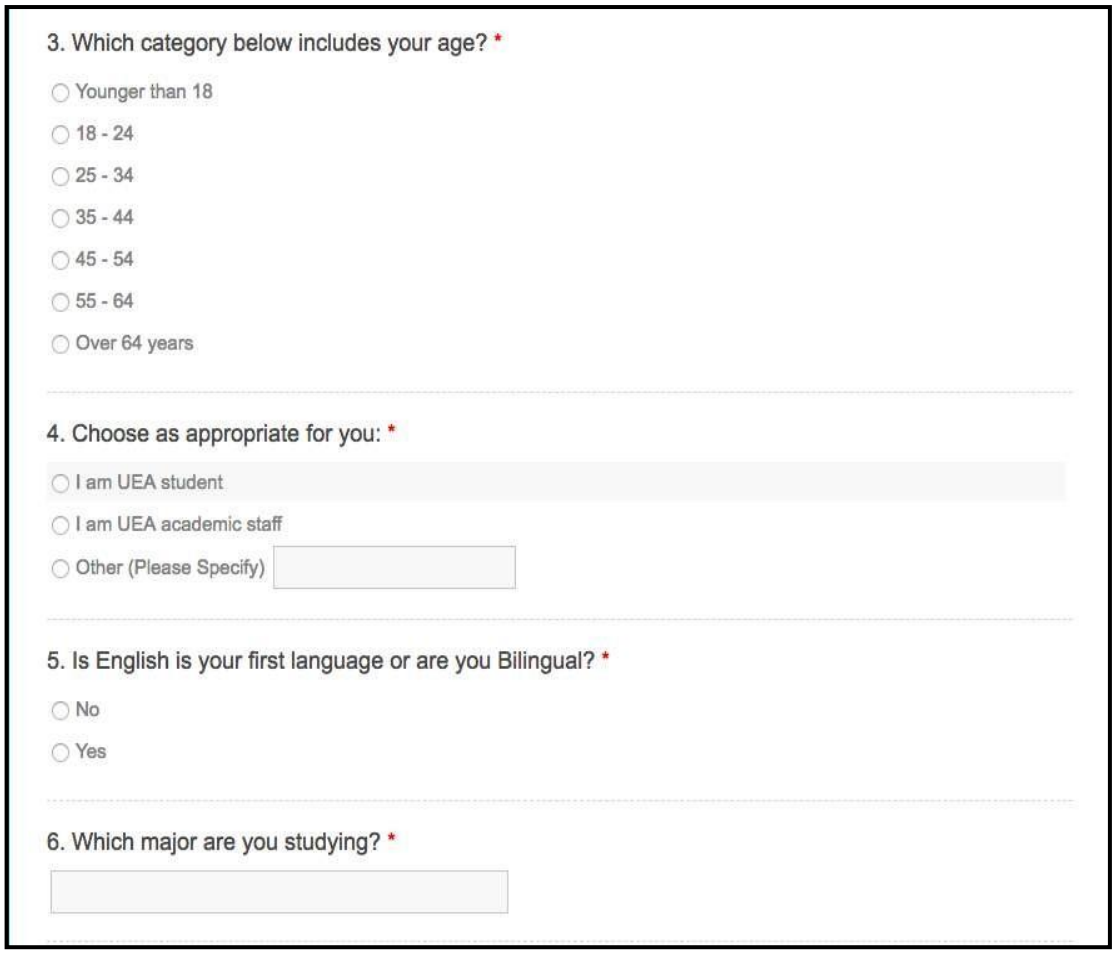


Figure A.CH6.5: Screening questionnaire (c)

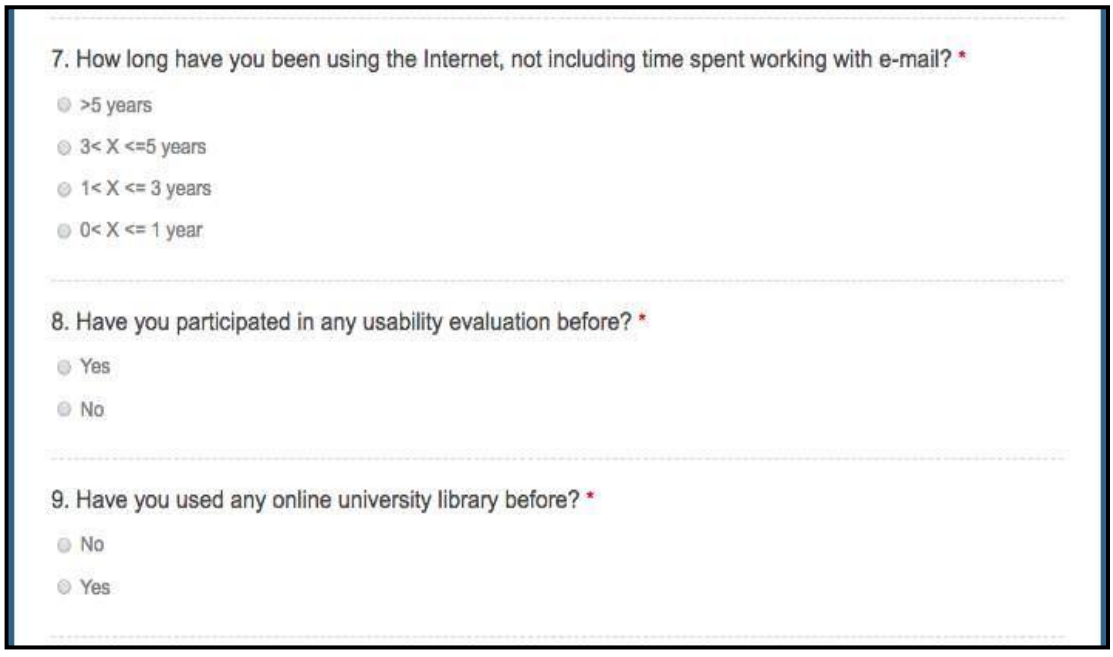


Figure A.CH6.6: Screening questionnaire (d)

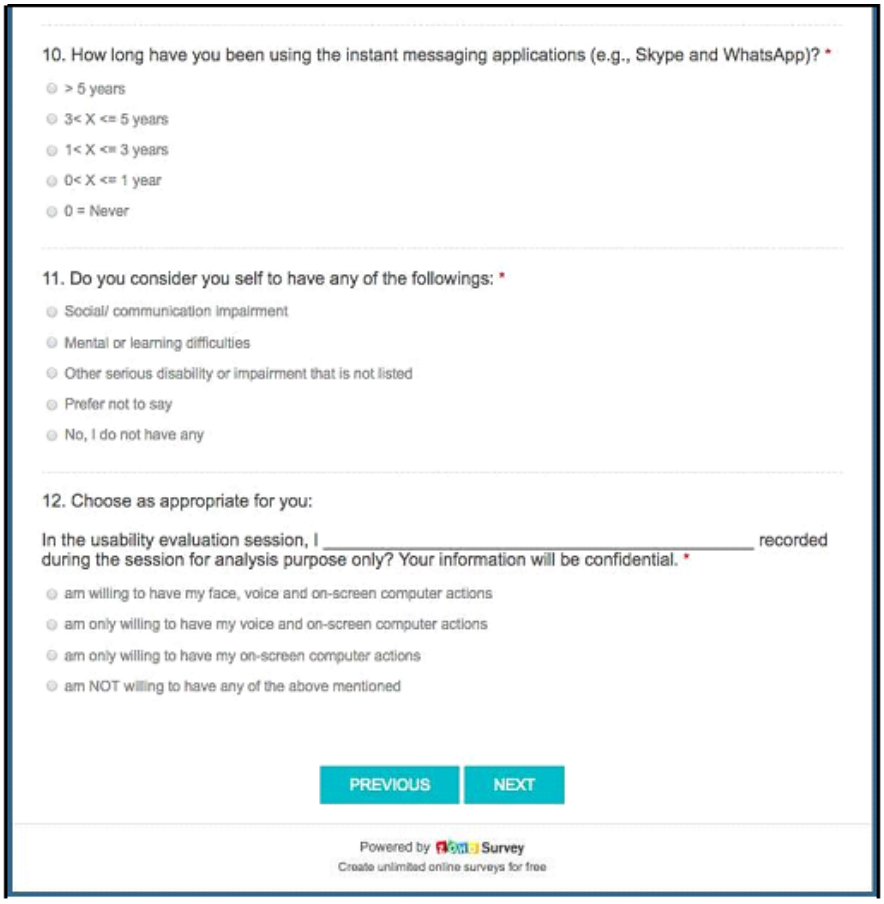


Figure A.CH6.5: Screening questionnaire (e)

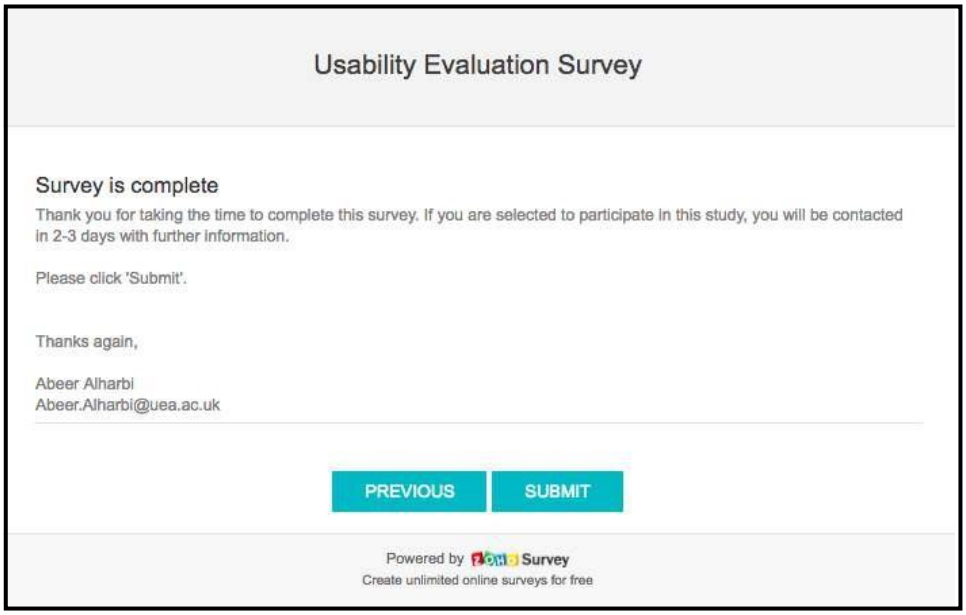


Figure A.CH6.6: Screening questionnaire (f)

A.CH6.4 Experimental Tasks

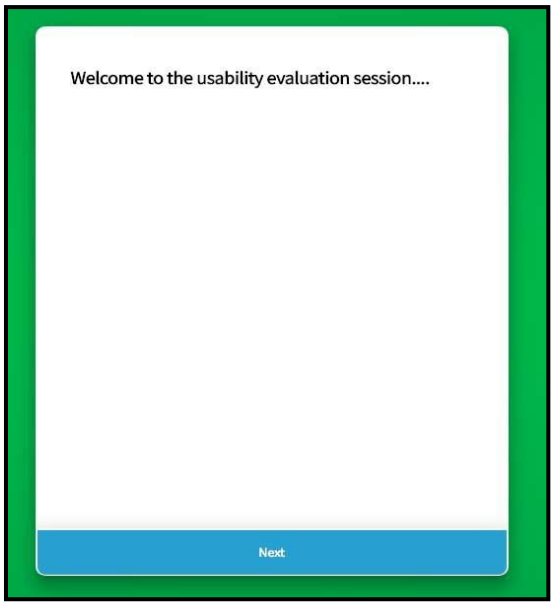


Figure A.CH6.7: Experimental tasks (welcome page)

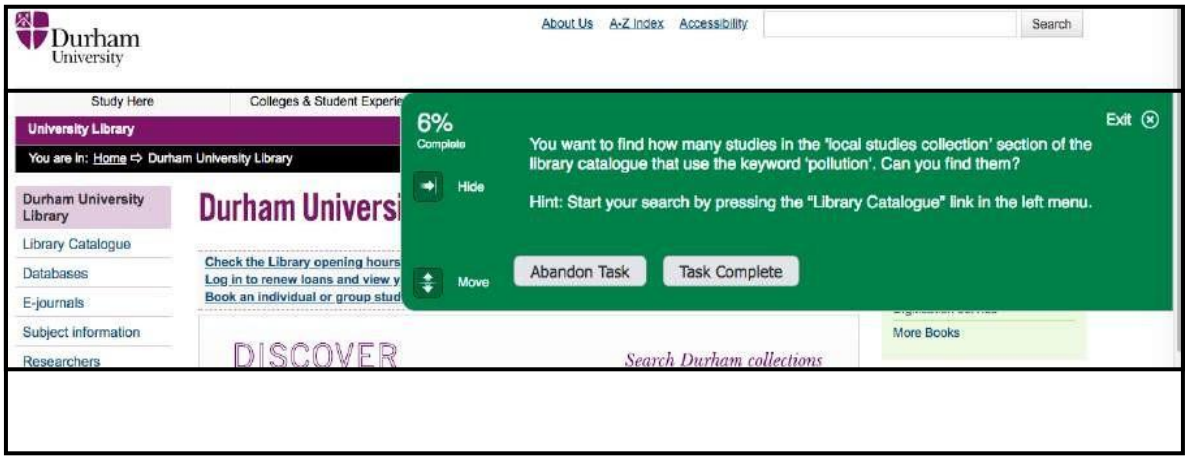
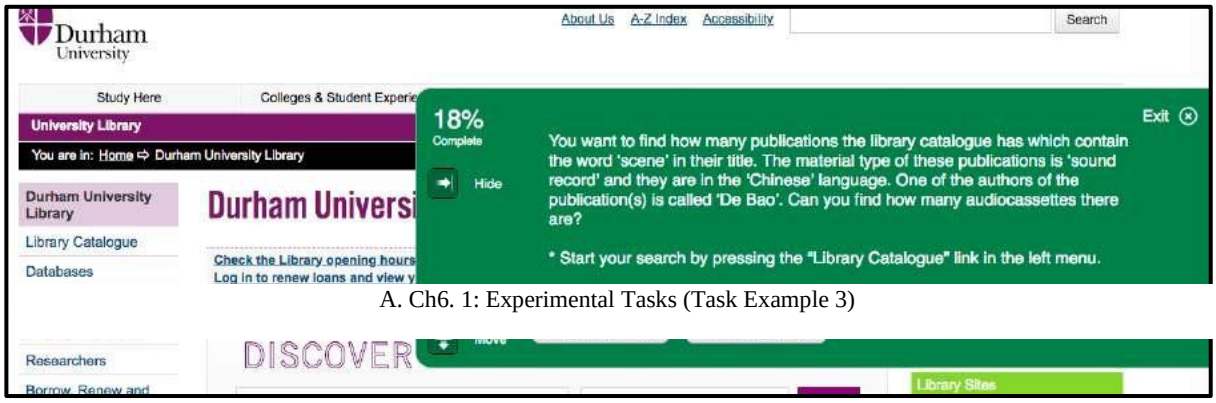


Figure A.CH6.8: Experimental tasks (Task example 1)

Figure A.CH6.9: Experimental tasks (Task example 2)



A. Ch6. 1: Experimental Tasks (Task Example 3)

Figure A.CH6.10: Experimental tasks (Task example 3)

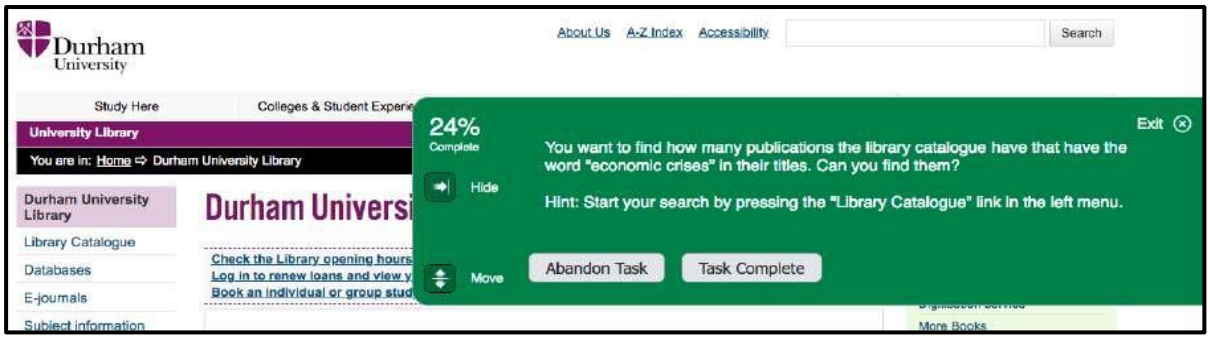


Figure A.CH6.11: Experimental tasks (Task example 4)

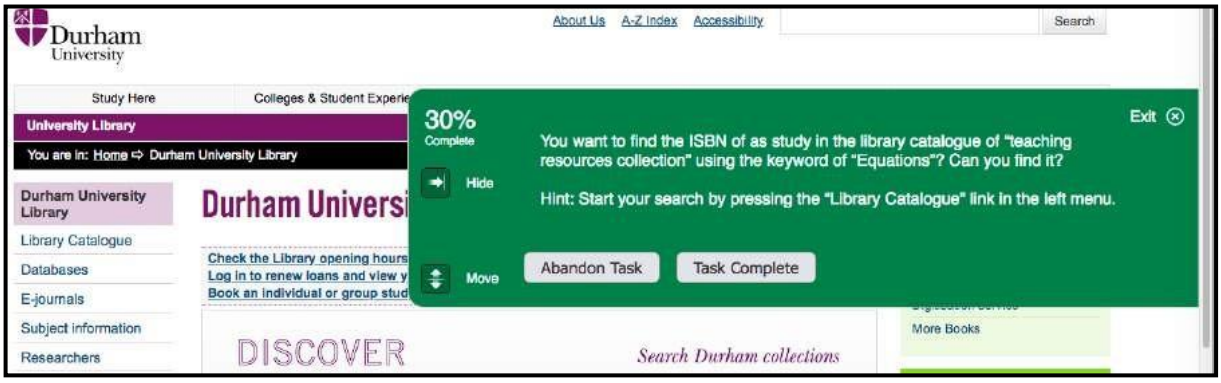


Figure A.CH6.13: Experimental tasks (Task example 5)

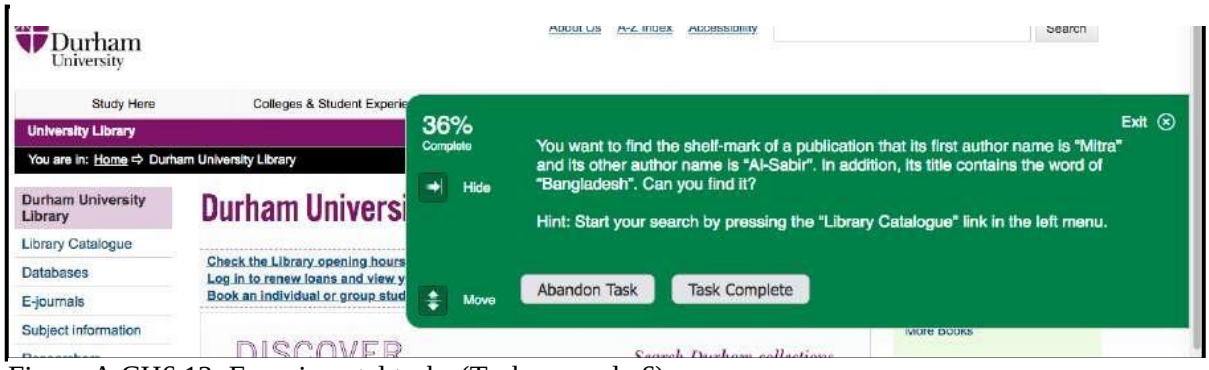


Figure A.CH6.12: Experimental tasks (Task example 6)

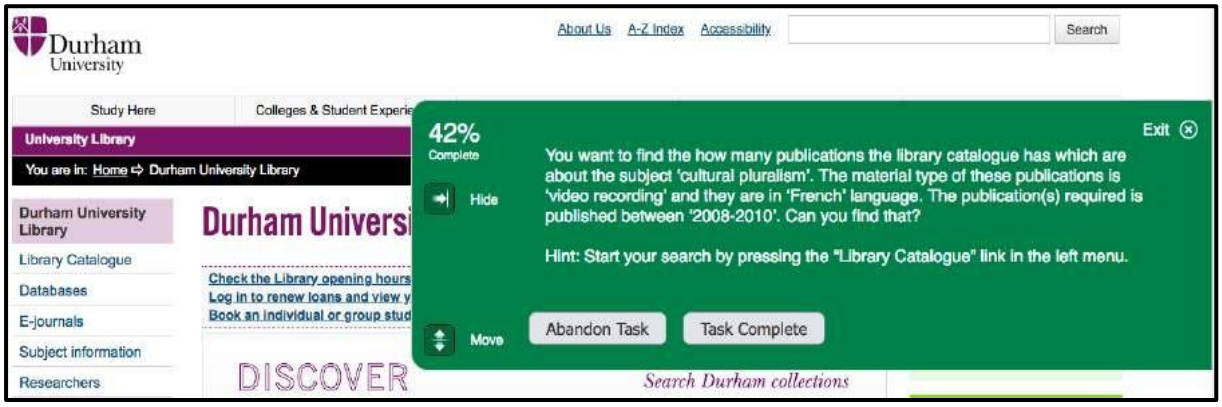


Figure A.CH6.17: Experimental tasks (Task example 7)

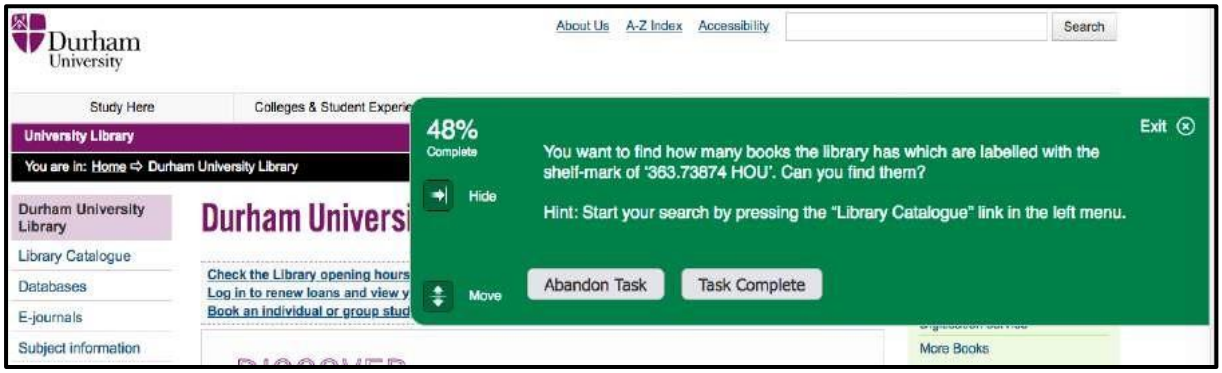


Figure A.CH6.16: Experimental tasks (Task example 8)

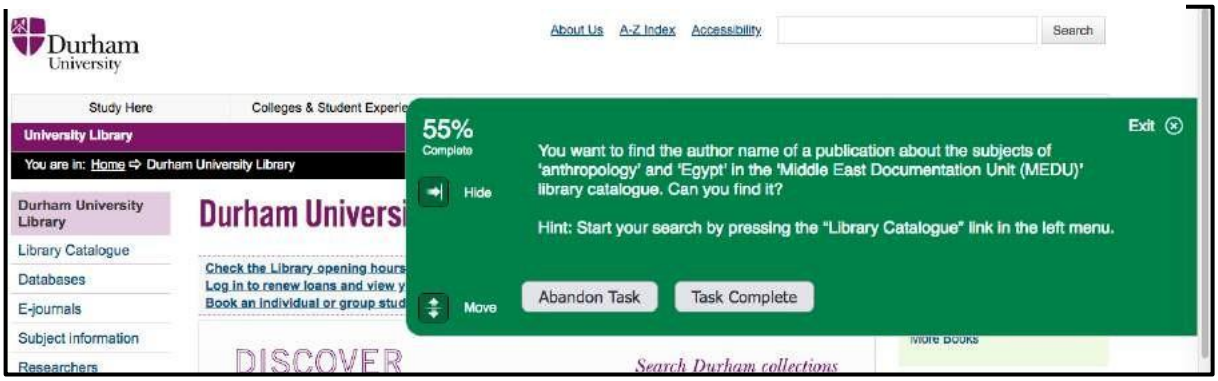


Figure A.CH6.15: Experimental tasks (Task example 9)

Figure A.CH6.14: Experimental tasks (Task example 10)

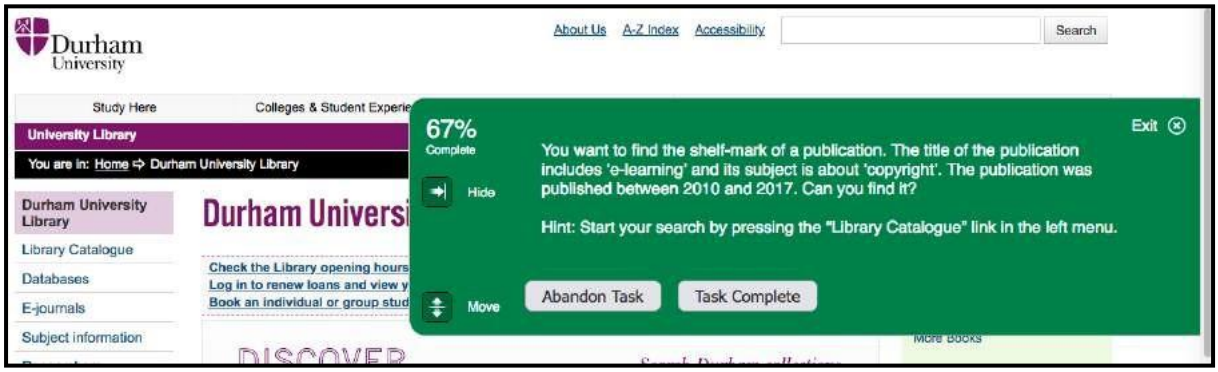


Figure A.CH6.18: Experimental tasks (Task example 11)

Figure A.CH6.21: Experimental tasks (Task example 12)

Figure A.CH6.19: Experimental tasks (Task example 13)

Figure A.CH6.20: Experimental tasks (Task example 14)

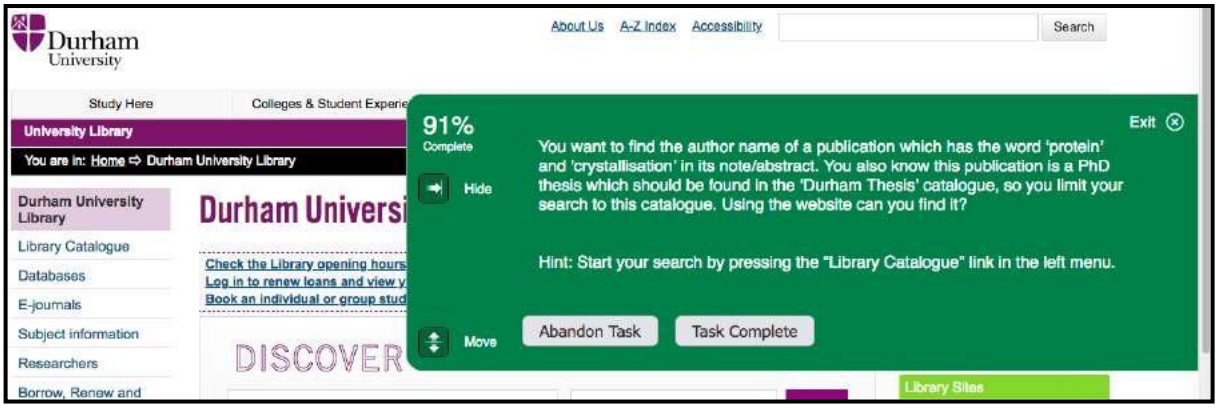


Figure A.CH6.22: Experimental tasks (Task example 15)

Figure A.CH6.23: Experimental tasks (Task example 16)

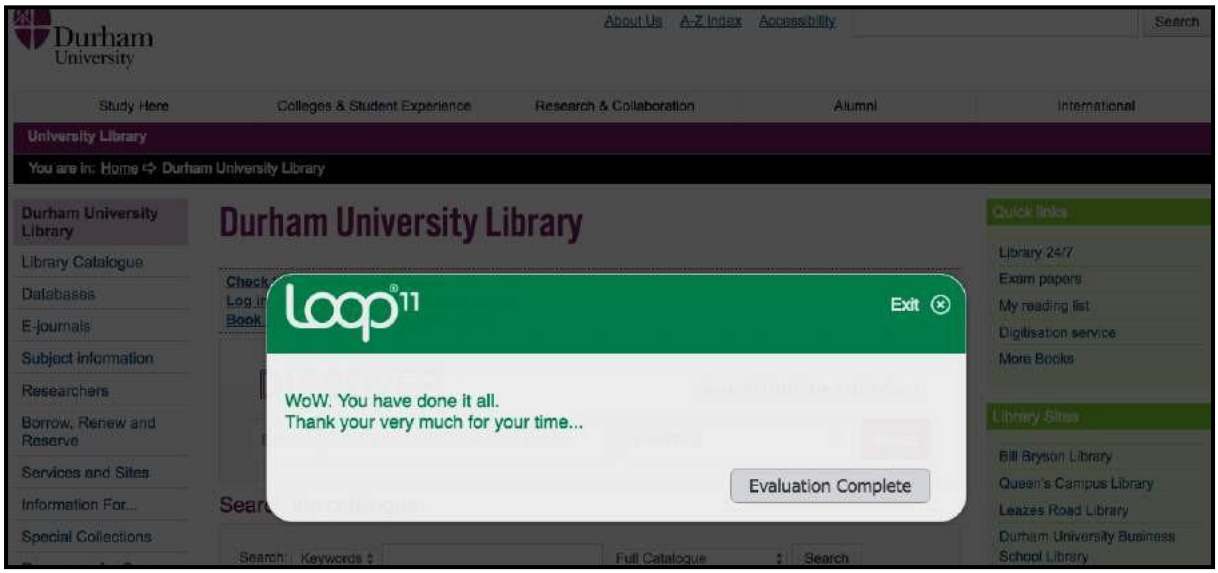


Figure A.CH6.24: Experimental tasks (Task example 17)

A.CH6.5 Incentive Receipt and Acknowledgment Form

Incentive receipt and Acknowledgment Form

I hereby acknowledge receipt of £10 for my participation in a research study run by Mrs. Abeer Alharbi.

Printed name:

_____

E-mail:

_____

Signature:

_____

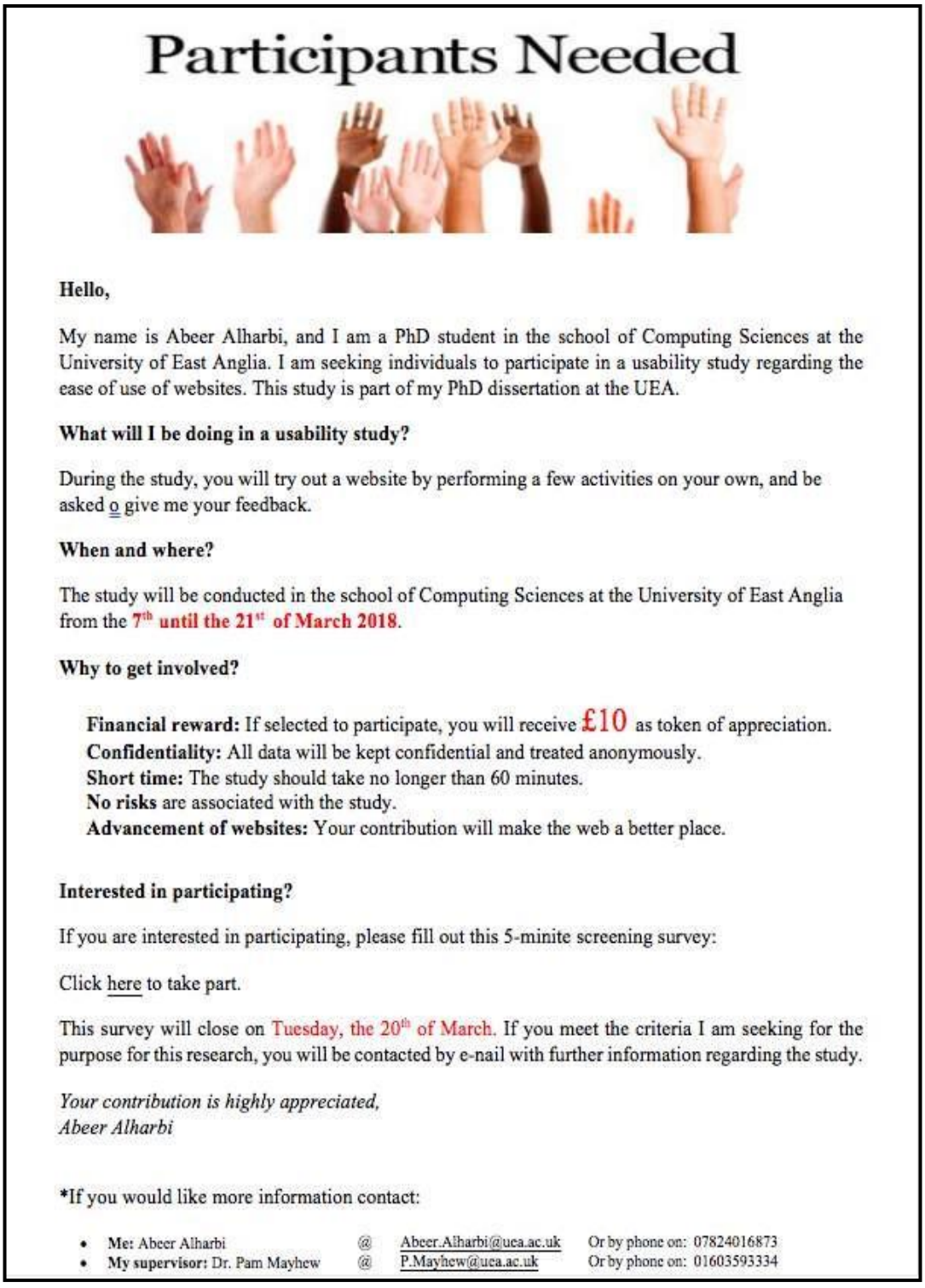
Date:

_____

Figure A.CH6.25: Transcript of the incentive receipt and acknowledgment form

A.CH6.5 Other Documents

A. CH6.5.1 Transcript Used for Advertisements



Participants Needed

Hello,

My name is Abeer Alharbi, and I am a PhD student in the school of Computing Sciences at the University of East Anglia. I am seeking individuals to participate in a usability study regarding the ease of use of websites. This study is part of my PhD dissertation at the UEA.

What will I be doing in a usability study?

During the study, you will try out a website by performing a few activities on your own, and be asked o give me your feedback.

When and where?

The study will be conducted in the school of Computing Sciences at the University of East Anglia from the **7th until the 21st of March 2018**.

Why to get involved?

Financial reward: If selected to participate, you will receive £10 as token of appreciation.
Confidentiality: All data will be kept confidential and treated anonymously.
Short time: The study should take no longer than 60 minutes.
No risks are associated with the study.
Advancement of websites: Your contribution will make the web a better place.

Interested in participating?

If you are interested in participating, please fill out this 5-minite screening survey:

Click here to take part.

This survey will close on Tuesday, the 20th of March. If you meet the criteria I am seeking for the purpose for this research, you will be contacted by e-nail with further information regarding the study.

Your contribution is highly appreciated,
Abeer Alharbi

*If you would like more information contact:

- **Me:** Abeer Alharbi @ Abeer.Alharbi@uea.ac.uk Or by phone on: 07824016873
- **My supervisor:** Dr. Pam Mayhew @ P.Mayhew@uea.ac.uk Or by phone on: 01603593334

Figure A.CH6.26: Transcript used for advertisements (“flyers”)

Figure A.CH6.27: Transcript used for advertisements ("posters")

A.CH6.5.2 Recruitment

Dear [participant name],

You are invited to participate in a usability study, where we will be evaluating the ease of use and user-friendliness of a website. You will be asked to use the website under evaluation, do a few tasks, and give your feedback. During the session, I will be capturing your screen and if you agree, record your face and voice; however, these recordings will be for research purposes only and will not be made public in any way. Please be assured that the purpose of this study is not to assess your skills or knowledge but rather to evaluate the usability of the website interface. The consent form will be detailed in the experiment.

The evaluation session will be held in room: _____ in the School of Computing Sciences at the University of East Anglia. The whole session is expected to take between 30-60 minutes. At the end of your session, you will receive £10 as a reward for your participation.

In order for me to reserve you place in the study schedule, please click on the link below and select the time that is most convenient for you to conduct the study. Please remember to type your full name in the required field, no one but I will have access to participants' names. It is extremely important that you keep your appointment with me. If for any reason you must reschedule, please contact me as soon as you know.

Link:

https://doodle.com/poll/ydupwbrn4yxgebsq

I will send you a reminder email two days before your session. Thank you for agreeing to participate in my study and for making the web *a better place*.

Sincerely,

Abeer Alharbi

Figure A.CH6.28: Transcript of the recruitment email

A.CH6.5.3 Reminder Email

Hello [participant name],

Thanks again for agreeing to participate in my usability study. This a friendly reminder that your session will be held in room: _____ in the School of Computing Sciences at the University of East Anglia on [date and time]. Please plan to arrive about 10 minutes before your scheduled session time. If you wear glasses while using the computer, please bring them with you to your session. Feel free to contact me with questions.

Many thanks,

Abeer Alharbi

Abeer.Alharbi@uea.ac.uk

Figure A.CH6.29: Transcript of the reminder email