THINKING ABOUT THINKING ALOUD:

An Investigation of Think-Aloud Methods in Usability Testing

Obead Alhadreti

A thesis submitted to the School of Computing Sciences of the University of East Anglia for the fulfilment of the degree of Doctor of Philosophy (PhD) in Computing Sciences

September 2016

ABSTRACT

In website design and engineering, the term "usability" describes how easy a website or interface is to use. As the Internet continues to grow exponentially, with millions of websites vying for users' attention, usability has become a critical factor determining whether a website will survive or fail. If websites are not sufficiently usable, users will simply abandon them in favour of alternatives that better cater to their needs. It is therefore crucial that designers employ effective evaluation methods in order to assess usability and improve user interface design.

One of the most widely used methods of evaluating the usability of websites is the Thinking Aloud protocol, wherein users are encouraged to verbalise their experiences, thoughts, actions, and feelings whilst interacting with the design. This provides direct insight into the cognitive processes employed by users—knowledge which can then inform strategies to improve usability. However, despite the common usage of Thinking Aloud protocol in the field, the specific think-aloud procedures employed vary widely among usability professionals.

The aim of this thesis is to investigate the utility and validity of the different variations of think-aloud usability testing methods. To this end, three empirical studies were conducted, using library websites, to compare the practical benefits of the various methods. The studies measured five points of comparison: overall task performance, the experiences of the test participants, the quantity and quality of usability problems discovered, the costs of employing the method in question, and the relationship between sample size and the number of problems detected.

Study One examined three classic think-aloud methods: concurrent think-aloud, retrospective think-aloud, and a hybrid method. The results revealed that the concurrent method outperformed both the retrospective method and the hybrid method in facilitating successful usability testing. It detected higher numbers of usability problems than the retrospective method, and produced output comparable to that of the hybrid method. The method received average to positive ratings from its users, and no reactivity (a potential issue wherein the act of verbalising the cognitive process alters that process) was observed. In addition, this method required much less time on the evaluator's part than did the other two methods, which involved double the testing and analysis time. Lastly, in terms of the relationship between the sample size and the number of problems discovered, the concurrent and the hybrid methods showed similar patterns, and both outperformed the retrospective method in this regard.

Study Two compared the performance of the classic concurrent think-aloud method with two variations on this method in which the evaluator plays a more active role—namely, the active intervention method and the speech-communication method. The results showed that these three methods enabled the identification of a similar number of usability problems and types, and showed similar patterns with regard to the relationship between the sample size and the number of problems discovered. However, the active intervention method was found to cause some reactivity, modifying participants' interactions with the interface, and negatively affecting their feelings towards the evaluator. The active intervention method also required much greater investment than did the other two methods, both in terms of evaluators' time, and, it was estimated, in financial terms.

Study Three compared the classic concurrent think-aloud method with the co-participation method, wherein a pair of participants work together to perform their tasks, and verbalise their processes as they interact with the interface and with one another. This study found no difference between the methods in terms of task performance. However, the co-participation method was evaluated more positively by users in comparison with the classic method. It led to the detection of more minor usability problems, and performed better in terms of the relationship between the sample size and the number of problems detected. The co-participation method was, however, found to require a greater investment of time on the part of the evaluator.

ACKNOWLEDGEMENTS

In the name of Allah, the most gracious, the most merciful

First and foremost, I give praise, honour and glory to Allah the Lord of the universe, without his bounty, grace and blessings this work would never have been accomplished.

I would like to express my heartfelt gratitude to my supervisor Dr. Pam Mayhew for her professional guidance and support during all stages of this research. I am most indebted to her patience and invaluable advice that inspired me to see things positively, and felt honoured with her confidence and trust in my ability. I would also like to extend my sincere thanks and appreciation to my co-supervisor Prof. Andy Day for all his constructive advice, support and suggestions.

My appreciation goes as well to Prof. Kasper Hornbæk from the University of Copenhagen and Prof. Andy Field from the University of Sussex, for their valuable consultations in the experiments design and analysis. Thanks should also be extended to all those who participated in the experiments for their valuable cooperation, thoughts, comments and suggestions. Without you, I would have had nothing to analyse and no thesis to write.

I am also grateful to the Saudi Government for giving me this opportunity to join the University of East Anglia, and gain unmatched educational experience in one of the best universities in the world. I am grateful for the great atmosphere and studying conditions that the School of Computing Sciences provided. Thanks to Ms. Sidney Brouet, Mr. Matthew Ladd, and Mr. Binoop Pulikkottil for making my work here easier. A further special gratitude goes to my dear friend Mr. Khalid Alosmani for his continuous support and encouragement.

Last but not least, I owe my loving thanks to my beloved wife Nourah, my son Wesam, and to my whole family, in Saudi Arabia, for their love, unremitting prayers, support and encouragement throughout my study in the UK.

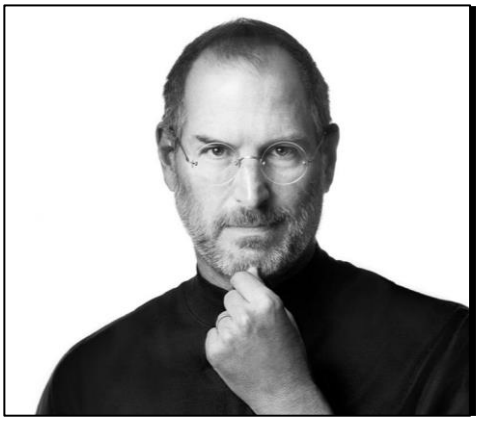
For those whom I have not acknowledged directly, I am very grateful to you all and wish you the success and pleasure you desire.

Obead Alhadreti
September 2016

DEDICATION

To my dear parents, wife, son, brothers and sisters

with sincere love and respect...



‘If a user is having a problem, it’s our problem’

Steve Jobs, Apple co-founder

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

RESEARCH INTRODUCTION

1.1 Overview

This chapter introduces the reader to the research, beginning by detailing the background and context that have informed it. The following section formulates the problem that this thesis will address. The chapter then discusses the aims and objectives of the research, and outlines the research questions. It also explains the motivations and significance of the research, as well as the methodology employed, and the phases of project. Finally, a brief description of each chapter of the thesis is provided.

1.2 Background

Usability is increasingly recognised as an important factor in the design and development of websites and web interfaces, offering multiple benefits for both development teams and end users. Several studies have demonstrated the benefits of a strong commitment to usability throughout the development life cycle of a product. These benefits include improvements in performance, safety, security, user productivity, and user satisfaction (ISO 13407, 1999). There are also significant cost- and time-saving effects—it has been estimated that the cost of correcting a problem after a product has been released can be as much as 100 times the cost of resolving it in the development phase (Aaron, 2005). The selection and employment of effective usability evaluation methods (UEMs) is therefore a crucial element of product development.

Over the last four decades, a number of different UEMs have been proposed (Scholtz, 2006). Amongst these methods, think-aloud (TA) methods, also known as TA protocols, are widely used (McDonald et al., 2012). The popularity of these methods stems mainly from their ability to offer insight into the thought processes and experiences of users interacting with a particular system during usability testing. The testing method has test participants work on a set of tasks, and asks them to verbalise their thoughts and task performance. Typically, the participants' verbalisations and behaviour are recorded, and a test evaluator is often present to observe and "read" the participants while working. As such, TA methods provide usability practitioners with verbal and visual indications of the usability of their systems (Clemmensen et al., 2009) (see Figure 1.1). The popularity of TA methods makes them an important area of research in usability testing. The next section briefly introduces the different types of TA methods.

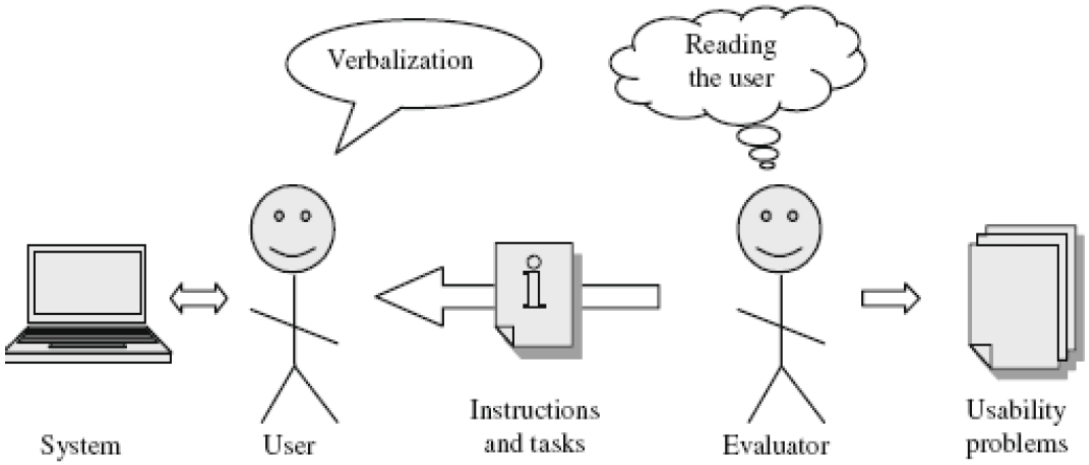


Figure 1.1: Think-aloud usability testing (Clemmensen et al., 2009)

1.2.1 Types of Think-Aloud Methods

TA methods were originally based on the theoretical framework developed by cognitive psychologists Ericsson and Simon (1980), and were introduced to the field of usability testing by Lewis and Rieman in 1982 (cited in Lewis and Rieman, 1993). According to Ericsson and Simon (1993), there are traditionally two basic types of TA methods: the *concurrent think-aloud (CTA)* method, in which participants think aloud at the same time as carrying out the experimental tasks; and the *retrospective think-aloud (RTA)* method, in which participants verbalise their thoughts after they have completed the experimental tasks.

The concurrent method provides "real-time" information during the participant's interaction with a system, which can make it easier to identify the areas of a system that cause problems for the user. However, there are three concerns. First, it might be an uncomfortable or unnatural experience, as people do not usually offer running commentaries whilst performing tasks. Second, the verbal reports are likely to be incomplete, since participants are expected to give priority to task solving, and may therefore forget to verbalise some thoughts. Third, the request to think aloud might interfere with and alter participants' thought processes, and may thus affect the ways in which they perform the experimental tasks—which can in turn affect the validity of the data obtained. This change is often referred to as *reactivity*. Reactivity may result in an improvement in participants' performances (e.g. by facilitating task completion or decreasing solution times), but it may also act as an impediment to performance (by

inhibiting task completion or increasing solution times). For usability testers, reactivity poses a problem: in cases where it enhances user performance, evaluators may fail to detect usability problems, or may assign unhelpful severity assessments. In the opposite case, where reactivity causes a decline in performance, evaluators risk identifying and, potentially, fixing problems that prove to be false positives (Zhao et al., 2012).

By contrast, the retrospective method does not interfere with participants' thought processes, but has been criticised for its reliance on memory, and the subsequent possibility of post-task rationalisations (Van den Haak et al., 2004). Ericsson and Simon (1993) advocate the use of concurrent and retrospective methods in tandem (referred to as the *hybrid (HB)* method in this thesis). This, they assert, offers a means of enriching the collected verbal data, and of strengthening the validity and reliability of verbal protocols, through the triangulation of concurrent and retrospective data. However, within usability testing, the hybrid method has received very little attention (McDonald et al., 2012). Indeed, in usability testing research, the concurrent and retrospective TA approaches are typically compared rather than combined (e.g., Peute et al., 2010; Ohnemus and Biers, 1993).

In TA studies, participant verbalisations offer valuable feedback on the product being tested. Ericsson and Simon (1993) argue that, in tests utilising the CTA method, verbalisations can only be considered valid if they represent directly accessible information contained in the participant's "working" or short-term memory (STM). Such verbalisations do not alter the sequence of information comprehended by participants, and so do not affect the tasks that participants perform during TA sessions. Conversely, any verbalisation that requires additional processing through reflection or elaboration, causing the flow of STM content to change during the TA process, is considered invalid. Ericsson and Simon, therefore, advise against the evaluator prompting or questioning the participant, since participants' verbalisations can be affected by interventions.

A slight exception regarding the validity of post-task verbalisation is made in the case of the RTA method. Since participants in RTA tests begin verbalising only after completing their tasks, they cannot verbalise information directly from the STM, but instead have to retrieve this from their long-term memory (LTM). Ericsson and Simon (1993) claim that,

if verbalisation takes place soon after task completion, and without any intervention from the evaluator, retrospective verbal reports can be regarded as valid data.

In the classic Ericsson and Simon model, therefore, interaction between participant and evaluator is regarded as a potential risk to the validity of data; and so usability practitioners should not interfere with participants during TA sessions, with the exception of reminding them to think aloud if participants fall silent for a period of 15 seconds. However, evidence gathered from field studies suggests that usability professionals often ignore the recommendations from Ericsson and Simon, choosing to adopt a more relaxed approach. These practitioners often intervene actively in the CTA process, exploring and questioning participants' reported experiences in the hope of extracting maximum utility from the data (McDonald et al., 2012; Shi, 2008; Nørgaard and Hornbaek, 2006; Boren and Ramey, 2000). This method is referred to as the *Active Intervention (AI)* method in this thesis. By intervening in this way, practitioners risk compromising the validity of the CTA test; additionally, there is no empirical evidence supporting the assumption that such interventions enhance the utility of the data collected.

The difference between classic CTA and the actual practices of usability professionals has led some researchers to question whether another approach to TA testing might be more effective. Boren and Ramey (2000) have proposed a theoretical alternative to the traditional protocol—referred to here as the *Speech Communication* (SC) method—where the evaluator takes on an "active listening" role. This is achieved through the use of acknowledgment phrases, which indicate to the participant that they are being heard and understood: that the evaluator is paying attention and is absorbed in the communication act. Aside from these affirmative phrases, no questions are asked, and no conversation is made. Boren and Ramey present their model as a compromise between the AI approach, which may risk skewing the validity of collected data, and the traditional CTA technique which requires the evaluator to listen passively, which some usability professionals (and participants) may find inadequate, uncomfortable, or unrealistic.

Another increasingly common variation of the TA methods outlined above is the *co-participation* (CP) method, also known as the *team TA* or *constructive interaction* method, wherein participants interact, not with the test evaluator, but with a second participant. In CP tests, a pair of participants work together to perform their tasks and engage in

verbalising as they interact (Adebesin et al., 2009). Though it is used less often than the single-user methods, this method is rapidly becoming more popular (McDonald et al., 2012). The main advantage of this method is that the test sessions are more natural than those utilising standard single-user TA tests, since people are more used to verbalising their thoughts when they are trying to solve a problem together. However, using two people for each test increases the cost of testing, and can make it difficult to recruit a sufficient number of test participants (Als et al., 2005).

The following section discusses the current state of TA research, and identifies gaps in the existing body of knowledge.

1.3 Research Problem

Despite the fact that there have been some efforts to study TA methods especially relating to the CTA method, so far, the knowledge of the contribution of TA methods to usability testing is inconclusive and incomplete. Indeed, usability testing research has been criticised as being problematic and in a state of crisis (Woolrych et al., 2011). This lack of understanding can be attributed to five main factors.

The first of these is that the research on usability testing methods is often of dubious quality (Hornbæk, 2010). If the literature is explored, it is often found that many studies do not use rigorous experimental designs (Gray and Salzman, 1998), fail to include a sufficient number of participants (Barkhuus and Rode, 2007), and/or fail to perform adequate statistical testing (Cairns, 2007).

The second factor is the lack of a thorough and holistic assessment of TA methods. TA methods have been evaluated based on a range of criteria, including usability problem identification (Peute et al., 2010), task performance metrics (Olmsted-Hawala et al., 2010; Van den Haak et al., 2004), participants' testing experiences (Zhao and McDonald, 2010), the cost of employing methods (Als et al., 2005), and the number of test participants needed to find a sufficient number of usability problems (Nielsen, 2000). However, no existing research addresses all of these criteria in a single study. The failure of previous studies to combine evaluation criteria has resulted in conflicting findings and an incomplete

collection of knowledge. The present research argues that a holistic assessment is essential to attain a thorough understanding of the contribution of TA methods to usability testing.

The third factor relates to the narrow focus on the number of problems detected. The majority of studies tend to use this as the only indicator for measuring the utility of a method (Hornbaek, 2010). This method works on the basis that all problems are of equal importance. There is often, in practice, a great deal of variation between problems: their seriousness, their types, and their value for future product optimisation. One of the main tasks of usability practitioners is to identify and prioritise problems. It is therefore vital that research in this area moves beyond counting problems and starts to closely examine the type and criticality of problems detected during testing (Hornbaek, 2010; Wixon, 2003).

The fourth factor relates to the "evaluator effect", defined as the extent to which "multiple evaluators evaluating the same interface with the same usability evaluation method detect markedly different sets of problems" (Hertzum and Jacobsen, 2001, p. 421). Research has shown that this effect can influence the reliability of the data collected (Howarth et aI., 2009; Law and Hvannberg, 2008; Capra, 2006). To arrive at reliable data on usability problems, it is necessary to control the evaluator effect by applying a detailed and structured approach for usability problem extraction. The majority of usability testing studies do not consider or discuss this factor (Hornbaek, 2010; Hornbaek and Frøkjeer, 2008).

The fifth factor concerns the gap that still exists between theoretical research into testing methods and usability testing as practiced in the field. As mentioned in section 1.2.1, field studies have noted that evaluators often tend to interact with participants, despite Ericsson and Simon's (1993) strong recommendation against this. Despite this, very few studies have investigated the utility of the more relaxed approaches. In a similar vein, there has been limited research into the CP method in the context of website usability evaluation, though the method is becoming increasingly common among professionals in the field. Finally, although Ericsson and Simon (1993) suggest collecting both concurrent and retrospective verbal protocols in order to obtain rich data, this hybrid approach has been discussed only rarely (McDonald et al., 2012).

It is clear from the above that many aspects of TA protocols as usability tools—particularly their validity and utility—deserve more methodological attention, and that there is still work to do before a deep understanding of the effects of different variations in TA protocols can be reached (Lewis, 2014).

1.4 Research Aim and Objectives

The broad aim of this research is to investigate the use of the different variations of TA methods in the context of website usability testing. These methods comprise the classic TA methods (the concurrent, the retrospective, and the hybrid methods), the relaxed TA methods (the active intervention, and the speech communication methods), and the co-participation method. The research aims to gain a substantial insight into the validity and utility of these methods, with a view to contributing to the existing body of knowledge regarding TA protocols. This will help usability practitioners to make more informed decisions about which TA variant to use in particular contexts. The methods selected for this research are either classical methods, or are commonly employed by usability practitioners (McDonald et al., 2012; Olmsted-Hawala et al., 2010).

The specific measurable objectives that must be achieved in order to accomplish the aim of the research are as follows:

1. To explore the current and relevant literature on usability testing and TA protocols. A solid understanding of the literature is necessary in order to identify gaps in the body of knowledge, and where improvements and contributions can be made.
2. To effectively and thoroughly plan a series of empirical studies which endeavour to meet the aim of the project.
3. To successfully carry out the planned studies to a high standard, producing conclusive results.
4. To analyse, scrutinise and compare the results of the TA methods investigated in order to evaluate each method's relative performance.
5. To discuss the findings and draw conclusions in terms of the research questions.
6. To provide a set of recommendations for the benefits of future researchers, as well as for usability practitioners and engineers considering TA methods for evaluating the usability of websites.

The extent to which these objectives have or have not been satisfied by the work contained in this thesis will be discussed in Chapter 8. Having extensively reviewed the existing literature on TA usability testing methods, the researcher can claim that this research is unique in its large-scale, holistic and systematic investigation of the use of the selected TA methods in usability testing.

1.5 Research Questions

This PhD research endeavours to address the following research questions:

Research Question 1 (RQ1): Are there discrepancies between think-aloud methods with regard to participants' task performances?

Research Question 2 (RQ2): Are there discrepancies between think-aloud methods with regard to participants' testing experiences?

Research Question 3 (RQ3): Are there discrepancies between think-aloud methods with regard to the quantity and quality of usability problems they detect?

Research Question 4 (RQ4): Are there discrepancies between think-aloud methods with regard to the cost of employing the methods?

Research Question 5 (RQ5): Are there discrepancies between think-aloud methods with regard to the relationship between sample size and number of problems detected?

The first research question examines the effect of each TA method on participants' task performance by looking at three criteria: the extent to which participants are successful in completing their tasks, the time they take to complete those tasks, and their navigational behaviour to determine whether the methods induce reactivity.

The second question investigates the ecological validity of the TA variations under study. Ecological validity is concerned with the extent to which test participants are able to interact with a system as they would in their natural environment. It is important for usability evaluators to ensure this type of validity, as test participants who feel stressed or uncomfortable about participating in a usability evaluation might fail to report a number

of usability problems that they would normally have noticed outside a test situation (Van den Haak et al., 2004). The risk of stress or discomfort on the part of participants is fairly high, as the settings in which TA tests are conducted usually differ from environments in which people would normally work with a system; such tests often involve a usability lab equipped with various tools to record participants' performances, as well as a test evaluator who observes participants as they perform tasks (Clemmensen et al., 2009) (see Figure 1.1). Placing participants within this environment could threaten the ecological validity of TA protocols and consequently affect the application of these methods.

The third research question does not require much justification, as comparing the number of problems identified by different UEMs has been described as a key measure in investigating the utility of UEMs (Molich and Dumas, 2008). To gain additional insight into the utility of the TA methods under investigation, the nature of the problems identified are also considered.

The fourth research question regarding the cost of employing the methods pays particular attention to the relative cost-effectiveness of the TA testing methods under investigation. If less time and money can be spent by evaluators on conducting and analysing tests whose outcomes are as satisfactory as those tests that require more time and money, then the former can be considered more cost-effective (Martin et al., 2014).

The final research question focuses on the relationship between sample size and the number of problems detected. Usability testers generally opt for five participants (Nielsen, 2000), but it remains highly questionable whether this number is sufficient (Lindgaard and Chattratichart, 2007; Molich et al., 2004). This research question seeks to investigate whether sample sizes work differently for the different TA methods under investigation.

1.6 Research Significance

The rapid growth of the World Wide Web, the significant increases in the number of people using websites, and the heavy investment from businesses into web-based systems all attest to the importance of improving the efficiency of website usability testing (Alshamari and Mayhew, 2008). As shown in section 1.3, there are many aspects of the use of TA variants within the context of website usability testing which deserve more attention.

Each step of this research project is designed to contribute to the advancement of knowledge regarding TA methods, which in turn will improve usability testing. Firstly, this research conducts an extensive review of current literature, providing a comprehensive analysis of the efforts of pioneers in this field such as Ericsson and Simon, Boren and Ramey, Hornbaek, McDonald, Nielsen, Wixon, and Van den Haak. It then identifies and investigates the most common variants of TA methods applied in usability research and practice. Notably, this study is the first to undertake a thorough and holistic examination of the influence of a range of TA protocols on the results of usability testing. This is achieved through a set of carefully designed and thoroughly explicated studies. Another unique factor is the application of a structured and explicit usability problem extraction approach to control for the evaluator effect. This represents a step forward for research into TA methods. Finally, where previous research has been criticised for its narrow focus on problem counting, this research employs a richer and more robust assessment strategy, which considers both the quantity and the quality of problems detected. This approach will offer a more comprehensive view into the effectiveness of a method.

1.7 Research Phases

Given the study's focus on investigating different variants of TA methods and the fact that TA testing methods are typically applied in usability laboratory settings (Norman and Panizzi, 2006), an experimental method is used in this research. The following paragraphs provide a global overview of the design of the research.

This research consists of three empirical studies, each of which addresses all of the research questions (see Table 1.1). Study One (classic think-aloud study) examined three classic think-aloud methods: concurrent think-aloud, retrospective think-aloud, and a hybrid method. In accordance with Ericsson and Simon's (1993) guidelines, the role of the evaluator was strictly non-interactive: the evaluator only intervened to remind participants to think aloud if they stopped verbalising their thoughts during testing for a period of 15 seconds. 60 participants were recruited for this study, with 20 participants allocated to each testing method. The numbers of participants, numbers of tasks, laboratory used, test object, and evaluation criteria were the same for each group. Only the TA methods varied between groups, as this was the issue under study. The data was analysed using both quantitative

and qualitative techniques as well as descriptive and inferential statistical analysis. A more detailed discussion of this study can be found in Chapter 4.

Study Two (relaxed think-aloud study) compared the performance of the classic CTA method with two relaxed variations on this method—namely, the active intervention method and the speech-communication method. The study involved three groups, each consisting of 20 participants. As with the first study, all conditions were identical; only the TA method employed varied between groups. This study will be discussed more thoroughly in Chapter 5.

Study Three (co-participation study) compared the classic CTA method with the co-participation method. This study involved a group of 40 participants for the CP method (which requires 2 participants per test session), and the data from a group of 20 participants was reused from the second study. As in the first and second studies, conditions were identical for both groups except for the TA methods used. A more detailed discussion of this study can be found in Chapter 6.

Table 1.1:Research Phases

	Goal	**TA Methods**
Study One: Classic TA Study	To investigate the classic TA methods	CTA, RTA, and HB
Study Two: Relaxed TA Study	To investigate the relaxed TA methods	CTA, SC, and AI
Study Three: Co-participation Study	To investigate the CP method	CTA and CP

1.8 Structure of the Thesis

The rest of this thesis is divided into seven chapters: Literature Review, Research Methodology, Classic Think-aloud Study, Relaxed Think-aloud Study, Co-participation Study, Discussion, and Research Conclusions. A brief outline of the contents of these chapters is provided below.

Chapter Two, Literature Review, explores the concept of website usability. It looks at usability evaluation methods, with a particular focus on TA methods, and the factors that may affect such evaluation. It also critically reviews previous studies of TA methods.

Chapter Three, Research Methodology, explores a number of possible research methodologies, and then presents the methodology used to address the research questions of this thesis. The chapter also outlines the factors influencing the design of the experiments, describes the data collection techniques employed in this research, and summarises the strategies considered for analysing the data.

Chapter Four, Classic Think-aloud Study, presents the first empirical study which, as mentioned in section 1.7, explores the impact of classic TA methods developed by Ericsson and Simon (1993) (CTA, RTA, and HB) on the outcome of usability tests. The chapter describes how the experiment was conducted, and reports the results obtained. It then provides a comparative analysis and discusses the findings of the study.

Chapter Five, Relaxed Think-aloud Study, presents the details of the second study, which compares the classic CTA method with the AI method and the SC method. The chapter discusses the approach taken to conduct the study, sets out the results obtained from the experiments, and discusses the main findings of the study.

Chapter Six, Co-participation Study, presents the details of the third study, which examines the effect of the CP method on the outcome of usability testing. It then reports, analyses, and discusses the findings.

Chapter Seven, Discussion, pulls together and highlights the main findings of the three studies, and engages in a critical discussion of these findings. This discussion will outline a number of recommendations and suggestions for usability practitioners with regard to TA testing methods.

Chapter Eight, Research Conclusions, summarises the research and its major findings, examining how and to what degree the aims and objectives of this research have been accomplished. It then details the main contributions of this research to the body of knowledge. Finally, it discusses the limitations of the research, and offers suggestions for future research into TA methods.

Figure 1.2 illustrates the connection between the chapters in this thesis and the appendices.

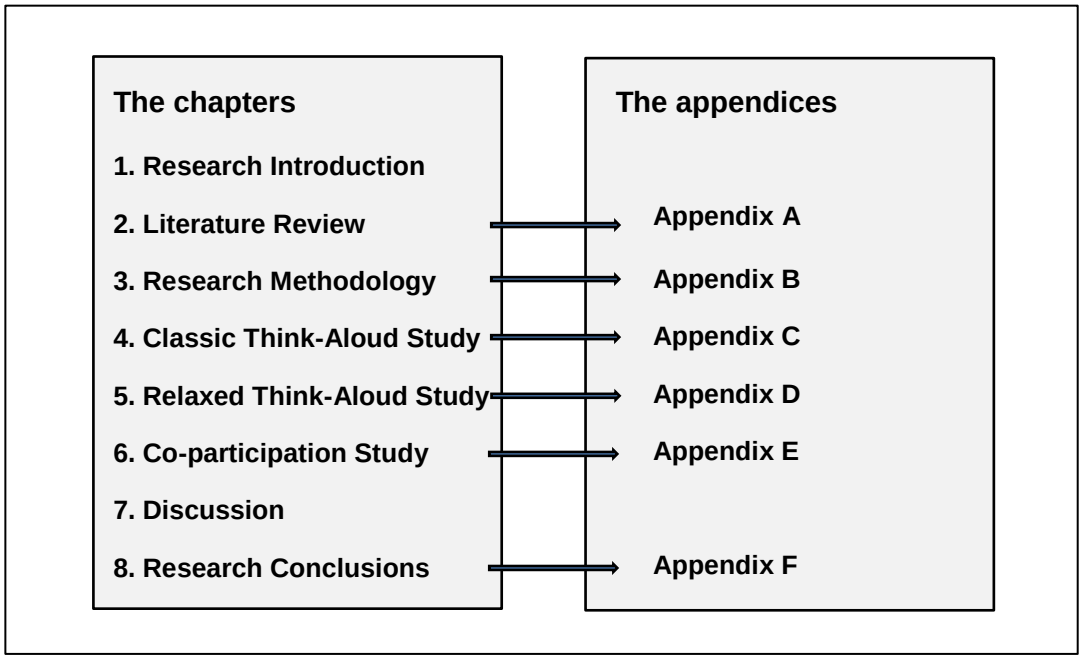


Figure 1.2: The connections between the thesis chapters and the appendices

2

LITERATURE REVIEW

2.1 Overview

Having introduced, in the previous chapter, the aims and objectives of the research, the thesis will now review the relevant literature. The chapter starts by defining website usability, highlighting its importance and looking at ways in which it can be achieved. This is followed by an overview of the various usability evaluation methods, a discussion of how a usability test can be conducted, and the factors that may affect the outcome of testing. The chapter then explores the different types of think-aloud (TA) methods, and looks at the previous comparative studies conducted on the methods. These studies are critiqued, and a knowledge gap is identified. Figure 2.1 below, provides a diagrammatic representation of this chapter.

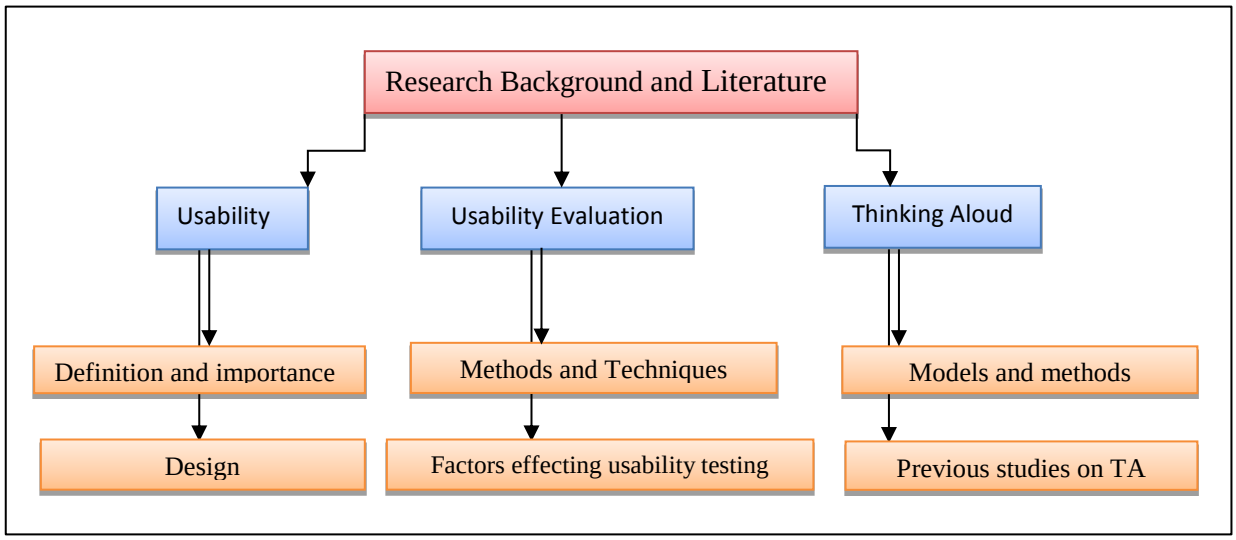


Figure 2.1: Diagrammatic representation of the literature review

2.2 Website Usability

Usability is a key concept in the field of human-computer interaction (HCI). HCI has been defined as a "discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" (Hewett et al., 1996, p.5). This section will discuss the various definitions of usability, and the ways in which it can be achieved.

2.2.1 Defining Usability

The term "usability" was coined in 1990s to replace the term "user-friendly". The existing body of literature defines the term in many different ways. The ISO standard 9241-11 (1996) and Nielsen's (1993a) definitions are probably the most widely used references (McNamara and Kirakowski, 2005). The International Standard ISO 9241-11 defines usability as "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". "Effectiveness" here refers to the accuracy and completeness with which users achieve specified goals. "Efficiency" means the amount of resources expended in relation to the product's effectiveness. "Satisfaction" means that users can complete their tasks without discomfort, and that they feel positive about using the product. Finally, the term "context" includes the users, their goals, the nature of the task(s), and the particular equipment, as well as the physical and social environments in which the product is used (see Figure 2.2). The usability of a product, then, is not simply an attribute of the product alone. Rather, it is an attribute of interaction with the product in a context of use (Karat, 1997). A product can therefore have very different levels of usability when used in different contexts. For this reason, the context should be clearly defined for design and evaluations (ISO 9241-11, 1996).

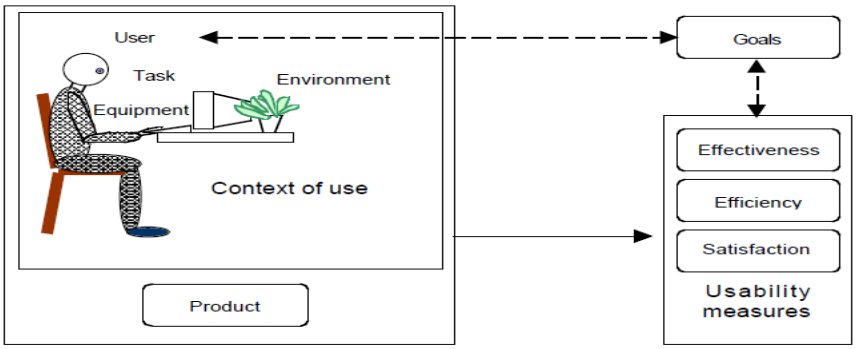


Figure 2.2: Usability framework according to ISO 9241-11 (1996)

Nielsen (1993a), on the other hand, defines usability as one of the main elements of a system's acceptability, which is the main question of whether the system is good enough to satisfy its end-users' needs and requirements (see Figure 2.3). In Nielsen's model, usability is subdivided into five main attributes: learnability, efficiency, memorability, error prevention, and satisfaction. "Learnability" means that new users should be able to easily learn to use the system. "Efficiency" means that the system should be efficient to

use once the user has achieved basic familiarity with it. "Memorability" means that the system should be easy to remember, even after a period of not using it. "Error prevention" means that the system should have a low error rate, and that users should be able to easily recover from possible errors. Finally, "satisfaction" means that the system should be pleasant to use.

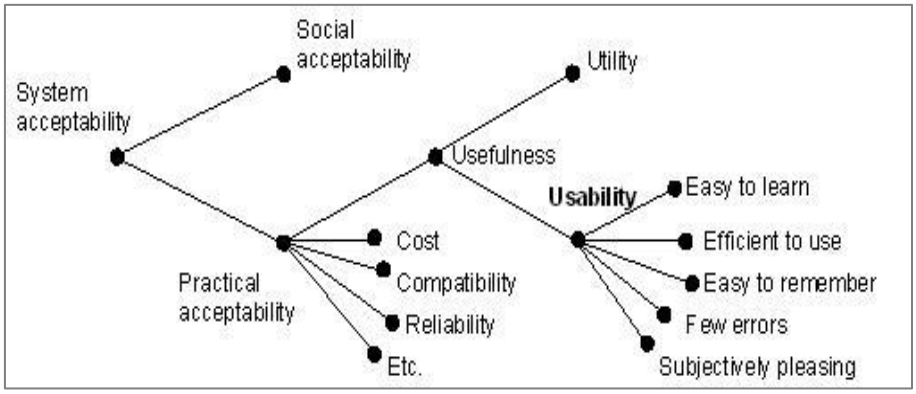


Figure 2.3: Usability as an aspect of system acceptability (Nielsen, 1993a)

It can be seen from Figure 2.3 that, in the Nielsen model, usability is a component part of system usefulness, which is in turn a component of practical acceptability, which is itself an element of system acceptability. Usability is therefore a major contributor to the perceived success of a system. For web based systems, usability is especially critical given that the web user population is expanding in age, expectations, information needs, tasks, and user abilities. Nielsen (1999, p.9) puts this very succinctly: "The web is the ultimate customer-empowering environment. He or she who clicks the mouse gets to decide everything. It is so easy to go elsewhere; all the competitors in the world are but a mouse click away". In other words, if websites are not sufficiently usable, users will simply abandon them in favour of alternatives that better cater to their needs. Despite the general recognition of the importance of usability for web based systems, it has been argued that many websites today still fail the most basic tests of usability (Choudrie et al, 2013). Appropriate website design and effective evaluation methods can help to ensure that websites are usable. The following section discusses the ways in which usability can be achieved.

2.2.2 Designing Usability

The International Standard 13407 (1999) provides a framework for designing usable interfaces. This is known as the *usability engineering lifecycle*, and is comprised of four activities that should take place during a system development project (Figure 2.4):

1. Understand and specify the context of use;
2. Specify the user and organisational requirements;
3. Produce design solutions;
4. Evaluate designs against requirements.

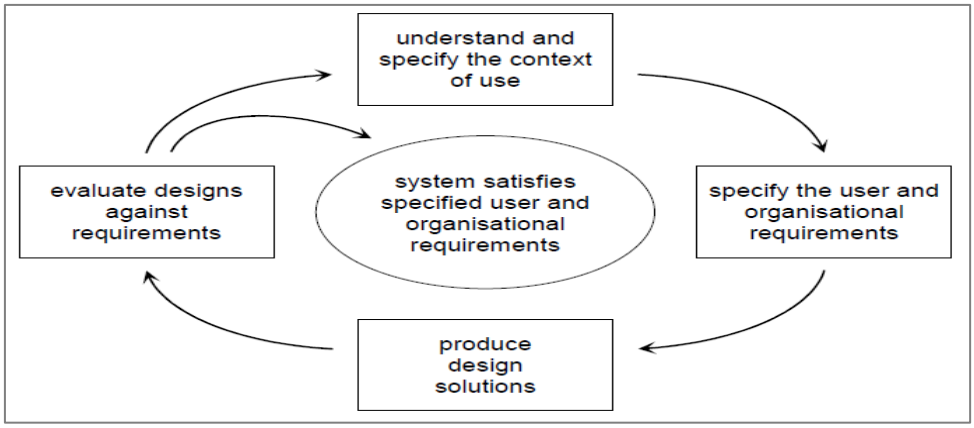


Figure 2.4: Usability Engineering Lifecycle according to ISO 13407 (1999)

Iteration is a key principle in usability engineering. This means that the cycle of analysis, design, implementation, and evaluation is continued until the iterative design has achieved its usability objectives (ISO 13407, 1999). This thesis focuses on the evaluation phase of this cycle. The next section provides an overview of the various usability evaluation methods that are available.

2.3 Usability Evaluation

In order to fully comprehend usability evaluation methods and their evaluation, one must first understand evaluation in the context of usability. Koutsabasis et al. (2007) define usability evaluation as the appraisal of a particular application's user interface, an interaction metaphor or method, or an input device, for the reason of ascertaining of determining its real or likely usability. Usability evaluation is required at several points during the design process. It is, however, important to start evaluation as early as possible,

because changes can become more expensive to implement as specific design and functionality decisions are made (Aaron, 2005).

Usability evaluation methods can be classified in numerous ways. The most common approach is to divide them into expert-based methods, model-based methods, and user-based methods (Dillon, 2001).

2.3.1 Expert-Based Methods

Expert-based methods (otherwise known as inspection methods) are a set of non-empirical methods that involve having experts assess the usability of an interface, predicting potential usability problems, and providing recommendations for improvement. Expert-based methods rely on the experience and knowledge of the experts, and so do not require extensive preparations or user involvement. As such, they can be swiftly and easily integrated into the development process. The two most commonly employed expert-based methods are *heuristic evaluation* and *cognitive walkthrough* (Scholtz, 2006). Heuristic evaluation, developed by Nielsen and Molich in 1990, involves inspectors checking whether the interface conforms to a set of guidelines or principles (Nielsen, 1995) (see Appendix A). Cognitive walkthrough, developed by Lewis in 1994, is based not on a set of guidelines but on a set of realistic task scenarios. By following these scenarios, experts attempt to discover the usability problems that users might encounter whilst working with the system (Nielsen, 1993a). The use of the verb "attempt" in this context is deliberate, as doubts are frequently raised regarding the validity of expert-based methods. It is often thought that, given their levels of expertise, the experts tasked with evaluating these systems are unlikely to detect real usability problems to a sufficient extent (Jong and Lentz, 1996).

2.3.2 Model-Based Methods

Model-based methods in usability evaluation are the least commonly used of the three methods. They stem from psychological research into human performance. The primary aim of adopting these methods is to predict certain aspects of user performance with an interface, such as total task time, or the difficulty of learning a task's sequence. A good example of a model-based method is the GOMS (Goals, Operators, Methods and Selection

Rules) model, which can be used to predict user behaviour, and to estimate the amount of time required for completing certain tasks. The detection of usability problems is not, however, the main purpose of methods such as GOMS (John and Kieras, 1996).

2.3.3 User-Based Methods

Many methods exist for conducting user-based evaluation, such as surveys, interviews, and focus groups. Another approach is to conduct usability testing using behavioural observation, such as think-aloud (TA) protocols (Lazar et al., 2010**)** , which are the focus of this research. Surveys, interviews and focus groups are methods which involve simply asking participants what they think of a particular test object. Surveys are usually conducted by means of a fixed set of questions, whereas interviews and focus groups are often semi-structured, consisting of either a face-to-face interview with a single participant or, in the case of focus groups, bringing together a small group of participants to discuss the benefits and drawbacks of a particular test object. Although all three methods are well established in the field of usability evaluation, as Nielsen (1993a) points out, they do have one main drawback: they only reveal what users *think* about a particular test object, not whether users can actually *work* with the object. As a result, behavioural observation is more widely used.

2.4 How to Conduct a Usability Test

Usability testing started to emerge in the early 1980's, and is most commonly used to test the usability of websites and software applications, particularly in the later stages of the development process (Rogers et al., 2011). Barnum (2011, p. 13) has defined usability testing as "the activity that focuses on observing users working with a product, performing tasks that are real and meaningful to them". Through testing, developers can gather information about how people interact with a system, and the problems that they encounter when doing so. The challenge for usability evaluators, however, is that they can see *what* a user is doing but not *why* they are doing it. The TA approach has been developed in response to this challenge. The general idea is for test participants to verbally express their intentions, actions, and frustrations whilst (or shortly after) working with an interactive system. From this data, the reasons behind their difficulties can be inferred and compared with the actual processes carried out by the participants (Rubin and Chisnell, 2008). The

usability practitioner then uses this information to identify problem areas of the system being assessed, and to offer recommendations for improvement. The main drawback to the TA method is that it can be time-consuming and expensive compared to expert-based or model-based evaluation methods (Molich and Dumas, 2008; Jeffries et al. 1991).

There are numerous handbooks on how to plan and conduct TA usability tests (e.g. Nielsen, 1993a; Dumas and Redish, 1999; Rubin and Chisnell, 2008). Dumas and Redish set out five specific requirements for usability testing:

1. A clear goal;
2. Real or representative users;
3. Real tasks;
4. Observation and recording; and
5. Analysing data and making suggestions for improvements.

According to Law and Hvanneberg (2004, p. 9), the primary goal of a usability test is to "derive a list of usability problems from evaluators' observations and analyses of users' verbal as well as non-verbal behaviour". Usability testing may also involve other metrics that seek to gauge usability by measuring performance and/or preference. Performance measures (e.g. time spent on tasks, or number of tasks completed successfully) indicate a user's level of capability with the system, whereas preference measures indicate how much the users enjoy using the system. Interestingly, a number of studies (Frøkjær et al., 2000; Hornbæk and Law, 2007; Nielsen and Levy, 1994) have found low correlations between user performance and user preference measures.

The involvement in testing of real or representative users who have not been involved in the design process is of critical importance (Holleran, 1991). In a TA test, the user is the participant who interacts with the system and verbalises his/her thoughts while doing the tasks. The tasks that the participant conducts and the instructions that the participant follows are set out by the evaluator. Apart from allocating tasks and giving instructions, the evaluator also needs to "read the user". This means that he/she has to observe the user's behaviour and listen to the user's verbalisations in order to understand the positive and negative aspects of the system (Nielsen, 1993a), and to achieve the goal of usability testing – the detection of usability problems (Hartson et al., 2001). Participants usually

work alone in usability tests, but testing in pairs can be more natural in some situations (Nielsen, 1993a).

Usability tests can be conducted practically anywhere: developments in the areas of computer networks and collaborative work tools mean that even remote testing is possible (Hartson et al. 1996). In general, however, usability tests are conducted either in specific usability laboratories, or in the field at the customer site. It is necessary to run a pilot test prior to the actual tests, in order to check the test tasks, instructions, and equipment. The pilot participant does not have to be from the target group, but should be somebody who is not part of the evaluation team. Dumas and Redish (1999) recommend that the pilot test is conducted two days before the actual tests are scheduled to take place, so that the preparations are finished but the test team still has enough time to make changes if needed.

After the test session, the evaluators analyse the data, diagnose the usability problems, and recommend changes to address the problems. It is important that evaluators list the problems in order of importance, so that developers can prioritise them accordingly (Dumas and Redish 1999). For example, problems can be classified according to their severity. The severity of a usability problem refers to the impact of the problem when it occurs. Several scales are available to rate these problems. Dumas and Redish (1999) suggest a four level scale with a clear reference to the impact on users' tasks:

- Level 1 problems prevent users from completing a task,
- Level 2 problems significantly slow down the user's performance and frustrate them,
- Level 3 problems have a minor effect on usability, and
- Level 4 problems point to potential enhancement in the future.

2.5 Factors Affecting Usability Testing

This section outlines factors that can potentially affect the results of a usability test. Andreas (2010) presents a framework of four factors that may influence usability testing. These are the test participants, the tasks provided, the system prototypes being tested, and the testing environment. Based on the literature review, this section will consider all four of these factors, along with two additional factors that Andreas does not mention. These are the so-called "evaluator effect"', and the effect of having to think aloud. This list of

factors is by no means all-inclusive, but it provides a solid basis for considering the consequences of various decisions made when planning a usability test.

2.5.1 Tasks

Task design has been shown to be a central issue to any usability evaluation method. Skov and Stage (2012) found that the quality and relevance of the test tasks significantly affected the number of problems detected. Usability testing tasks should, therefore, accurately represent the activities that real users would perform when using an application in order to achieve certain goals. Hansen (1991) recommends forming a group with representatives from the customer organisation to select the tasks. People from various parts of the organisation can offer different insights into the critical tasks, and participating in the design process can make them more supportive of the testing. Tasks can also be selected to test the use of specific but presumably problematic parts of the system. If less interesting functions are tested and the problematic functions are not covered, the whole process would have been a waste. As Munzner (2003, p.14) says:

> *"A study is not very interesting if it shows a nice result for a task that nobody will ever actually do, or a task much less common or important than some other task. You need to convince the reader that your tasks are a reasonable abstraction of the real-world tasks done by your target users*."

The tasks should be meaningful, and be presented to the participant in a logical order (Hansen, 1991). The tasks should also be independent from one other and should be presented to the participant one at a time. The instructions should clearly describe the goal of the task without telling the user how to achieve it. The task scenario should also be brief, and should use ordinary language rather than product or field-specific jargon (Dumas and Redish 1999).

2.5.2 Participant Effect

According to existing literature, there are two major influences that must be taken into account before selecting participants for testing: number of participants (sample size), and relevance of participants.

Taking into account human variation and the differences between individuals, it is clear that studying one participant would be insufficient to capture the majority of problems in an interface. The question of how many participants are sufficient, however, is a matter of some debate. Various studies have investigated the most effective sample sizes in TA usability testing (predominantly studying the concurrent TA method). Virzi (1992) was the first to investigate this issue. Based on three different experiments Virzi found that only five participants were necessary in order to capture 80% of the usability problems. Nielsen has also conducted a number of influential studies (Nielsen and Landauer, 1993; Nielsen, 1994; Nielsen, 2000). Nielsen and Landauer (1993) first found that they needed between four and nine users to find 80% of the usability problems. However, Nielsen's final recommendation was to plan for five participants to find 85% of the problems (Nielsen, 2000) (see Figure 2.5).

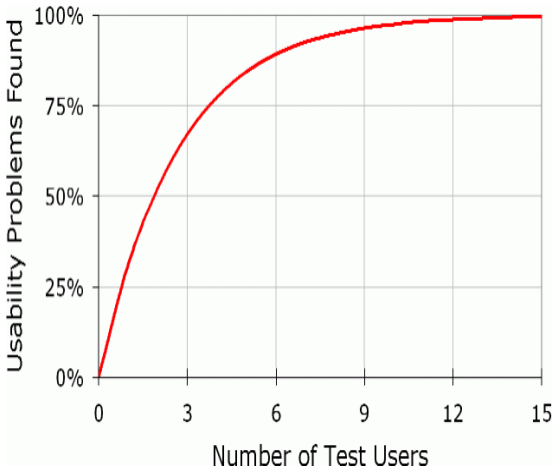


Figure 2.5: Curve showing relationship between problems found and number of users (Nielsen, 2000)

Due to the significance of the issue, the 2003 HCI conference hosted a panel officially named "The Magic Number 5", which discussed Nielsen's controversial claim. Its opponents argued that five participants are only sufficient when problems are relatively easy to find; and they emphasised the importance of other usability variables, such as task protocol, and the condition of the system in terms of interface design. They added that employing only a small number of users ignores the individual differences between them (Bevan et al. 2003). Several articles support these arguments by demonstrating that critical usability problems may be missed when a group of only five participants is involved. For instance, Bevan et al. (2003) conducted a study on four e-commerce websites and found

that five participants discovered only 35% of the usability problems. Another study by Molich et al. (2004) compared the performance of nine teams and found that the top-performing team found only 75% of the total usability problems identified by all the teams together. Lindgaard and Chattratichart (2007) conducted nine tests and compared the results of two teams, where team A consisted of six users and team B consisted of twelve. The analyses showed that the teams discovered 42% and 43%, of problems respectively.

With these conflicting results, the question of what sort of sample size is required in order to find a sufficient number (e.g. 85%) of usability problems certainly deserves more attention. The study of sample sizes is crucial: it can benefit usability evaluators by helping them cut the costs of their practice without compromising efficiency.

There is general agreement between usability researchers that, regardless of size, a test sample should be as representative as possible of the targeted users of the tested system. Relevant users are more likely to encounter relevant problems, which in turn will produce more relevant results. Possible criteria that can be used to define the test user sample include their level of experience with the Internet, website interface being evaluated, and usability evaluation (Sova and Nilesen, 2003).

2.5.3 Evaluator Effect

The "evaluator effect" refers to the observation that individual usability evaluators can identify substantially different sets of usability problems when analysing the same test sessions (Hertzum et al., 2014). A range of approaches have been taken when studying the variation in results between evaluators. These include studying the same video recordings from the same usability test sessions (Jacobsen et al., 1998; Vermeeren et al., 2003; Hertzum et al., 2014); and comparing the results of different groups evaluating the same system with the same goals and instructions (Molich and Dumas 2008). The study by Jacobsen et al. (1998) is one of the first to compare the usability problems derived by different evaluators from identical video recordings. In this study, all four evaluators were experienced in usability testing. The evaluators were asked to list and describe all the usability problems that they could detect from the video recordings, and to identify the ten most important problems to be fixed in the next release of the system. The results showed that each evaluator detected between 39% and 63% of the total number of problems; only

20% of the problems were detected by all the evaluators, and 46% were detected by only one evaluator. For the lists of the ten most severe problems, the evaluator effect was less drastic; nonetheless, no single problem was common to all the evaluators' lists (Jacobsen et al. 1998). Subsequent studies by Vermeeren et al. (2003) (which used 2 evaluators and 3 different tests), and by Hertzum et al. (2014) (19 evaluators analysing 1 case) lend support to the issue of the evaluator effect. In studies conducted by Molich et al. (2004), and Molich and Dumas (2008), the results of the various teams overlapped very little.

To ensure that the results of usability tests, particularly those in research studies, are reliable, it is preferable to collect usability problems after the fact, from video footage, than to note down the problems on-the-fly. This latter approach severely decreases the evaluator's ability to accurately record participant behaviour, as the participant does not stop working while the evaluator records problems. Hence, if two or more problems follow each other closely, only one of them might be detected and recorded. In addition, if the usability test has been conducted with concurrent note-taking as the only recording method, these notes form the sole basis for later analysis. The notes offer only a weak representation of the situation, since there has already been an element of selection or editing in terms of which aspects of the situation were recorded. Thus, the data to be interpreted is already based on an interpretation of a situation, which means that observation and analysis melt together when conducting on-the-fly usability tests (Jacobsen et al., 1998).

The evaluator effect can also be minimised through a common understanding of usability criteria, and by specifying defined scales of measurement. Often usability evaluators do not specify exactly what they are looking for, other than "usability problems". If a usability problem is not explicitly defined in concrete terms, it cannot be reliably identified. To further reduce the evaluator effect, it is also recommended that additional evaluators are involved in analysing the data (Barendregt, 2006).

2.5.4 System Prototypes

According to Rudd et al. (1996), prototypes can be classified into two broad groups: low-fidelity prototypes with limited functions that demonstrate the general look of the interface instead of its full operation; and high-fidelity prototypes that usually include complete

functionality and allow users to explore the system as if it was the final product. Low-fidelity prototypes are particularly valuable in the early phases of product development to help get a sense of what is required from the product (Rudd et al. 1996), whereas high-fidelity prototypes are useful in the later stages, when estimates of performance measures are needed (Virzi et al. 1996). Usability tests can be conducted both with low and high-fidelity prototypes as well as with finished products.

The effect of the prototype has been the focus of several studies. These studies have included comparisons between paper prototypes and interactive software simulations (Virzi, et al., 1996; Catani and Biers, 1998; Boothe et al., 2013), as well as comparisons between prototypes and the real physical products (Archer and Yuan, 1995; Sauer and Sonderegger 2009; Sauer et al. 2010). The first studies comparing the use of low- and high-fidelity prototypes show that both sorts of prototypes reveal substantially the same sets of usability problems. For example, in studies conducted by Virzi et al. (1996), low-fidelity prototypes consisting of paper cards representing the screen and keyboard in various actions, and a moderator simulating a voice response system, were compared with high-fidelity prototypes. The results showed that the prototypes revealed similar sets of usability problems, and even the proportions of test users detecting particular problems were remarkably similar (Virzi et al. 1996). The results of subsequent studies conducted by Catani and Biers (1998); Sauer and Sonderegger (2009); and Sauer et al. (2010), which utilised three different levels of prototype, support these findings. These studies all reported similar performance results and subjective evaluations between the different prototypes.

A study by Boothe et al. (2013) focused on the medium of the prototypes. The experiment uses the same user interface, presented to participants either as printed hard copies, or as a slideshow operated by a evaluator on a computer. In line with the findings by Virzi et al. (1996), the results showed that the medium of a prototype does not affect the probability of participants identifying usability problems. Boothe et al. did, however, find that the computer medium was more effective when it came to identifying severe problems. The subjective evaluation of the system's usability remained the same irrespective of prototype medium (Boothe et al. 2013).

2.5.5 Test Environment

As mentioned earlier, usability tests can be conducted anywhere. As Anna Wichansky (2000, p. 1004) has pointed out:

> *"Today, usability testing is being conducted in simulated homes, classrooms, cars and virtual reality environments. There are portable lab systems that can be carried to remote user sites to collect data, so usability engineers can go to their users if their users cannot come to them."*

A usability test can be conducted in a dedicated usability laboratory, or in the field, or in any setting in between these two extremes (Rubin and Chisnell, 2008). The real use context, with tasks emerging from the users' work, reveals problems that would be hard to detect in laboratory settings with predefined tasks. For example, McDonald et al. (2006) estimated from their data that about 2/3 of the problems identified in their study were related to the context of use instead of the evaluated system. Consequently, the real context of use, tasks emerging from the users, and a rich data set are considered to be the main advantages of the field methods. Disadvantages include the potential of being laborious, the greater time investment required, and problems in data analysis (Monahan et al. 2008). The customer site is familiar to the participants, making it easier for them to relax, but is more challenging for evaluators, as interruptions are hard to control, and the available equipment varies from site to site, or has to be brought along specially. Specific laboratories, on the other hand, offer dedicated equipment and a peaceful environment, but the participants must then be willing to travel to these laboratories. In addition, the artificial environment can produce unrealistic results. Nonetheless, testing in laboratories gives greater control of the variables critically affecting the level of usability, and the measurements obtained are more precise than in the field tests (Rubin and Chisnell, 2008).

2.5.6 Thinking Aloud Effect

As mentioned earlier, thinking aloud is a method to follow a user's plans, actions, and opinions. Verbalized plans are to help the evaluator to understand what the user is about to do, and why the user is clicking buttons or in other ways interacting with the system. Thinking aloud about user's preferences and opinions is, according to ISO/DIS 9241-11 (1996), an important aspect of usability and might lead to problem detection if users are

frustrated about certain parts of the interface. Nevertheless, the TA method is not a method without problems. At the present time, the use of the TA methods are at the centre of a debate (Woolrych et al., 2011). For instance, a number of researchers argue that thinking aloud while performing tasks affects the behaviour of participants in usability evaluations (e.g. Oostendorp and De Mul, 1999), while others claim thinking aloud does not affect user performance (e.g. Hertzum et aI., 2009). Although there is some evidence in support of these claims, the evidence is mixed.

Furthermore, previous research has revealed that the specific TA procedures employed vary widely among usability professionals and researchers. This has hindered the emergence of a coherent body of knowledge around TA methods. This lack of understanding explains why the validity and utility of the TA methods for usability evaluation is presently debatable.

The next section will discuss more thoroughly the theoretical background and different types of TA methods.

2.6 Think-Aloud Methods

2.6.1 History and Theoretical Background

Despite their increasing use within the context of usability testing, TA methods were originally developed within a relatively narrow niche in the field of cognitive psychology. John Watson (1920) was the first to report on using thinking aloud as he tried to learn more about the psychology of thinking (Fox et al., 2011). Duncker (1945; original German version 1935) was among the first researchers to utilise thinking aloud in empirical studies of mathematical problem solving in 1925-40. Later, the verbal reports produced by TA protocols also began to serve as a basis for discovering how people perform certain activities in many other fields: how they write (Hayes and Flower, 1983) or read (Ericsson, 1988); what a translation process looks like (Séguinot, 1996), et cetera. Most of the literature devoted to TA protocols is based, to a larger or smaller extent, on Ericsson and Simon (1980), whose influential work has almost single-handedly validated the use of verbal protocols as research data.

Thinking Aloud in Usability Testing

TA methods have been employed in usability testing for more than thirty years since their introduction to the field by Lewis and Mack in 1982 (cited in Lewis and Rieman, 1993), when the concurrent think-aloud (CTA) method was used to get insight into the users' mental processes as they learned to use new text processing systems. Studies by Jørgensen (1990) and Wright and Monk (1991) have shown that TA methods are highly effective for detecting usability problems in user interface design, especially if the designers conduct the usability tests themselves and so get direct feedback from the users. Since then, TA methods, have become the methods of choice for many usability practitioners (Kumar et al., 2008). In a survey of methods used by usability practitioners (about 75% of respondents) and researchers (about 25% of respondents) in Denmark, TA appeared to be the single most frequently applied method of evaluation (Clemmensen, 2002).

This should not come as a surprise—the TA methods are taught as part of the HCI curriculum at many universities around the world, and are described in many textbooks. The textbooks on usability evaluation published in the early 1990's (e.g. Nielsen 1993a) established TA methods as a central component of usability testing practice. The studies by Ericsson and Simon (1980, 1993) are sometimes cited as references for thinking aloud in usability testing (e.g. Nielsen, 1993a), but quite often the method is introduced without any references (e.g. Tullis and Albert, 2008; Dumas and Loring, 2008).

The next section provides a thorough overview of the different types of TA methods considered in this thesis, namely the classic TA, the relaxed TA, and the co-participation methods.

2.6.2 Classic Think-Aloud Methods

The classic TA methods are the methods described by Ericsson and Simon (1993): the *concurrent think-aloud method*, the *retrospective think-aloud method*, and the *hybrid method*.

2.6.2.1 Concurrent Think-Aloud Method

Concurrent think-aloud (CTA) requires participants to verbalise their actions and thought processes in real time, whilst they are completing the test tasks. This method is the most common TA variant in the field of usability testing (Nielsen, 1993a). Indeed, in an international survey conducted by McDonald et al. (2012), 98% of respondents had utilised CTA, and 89% rated it as the most frequently used approach (see Figure 2.6). CTA is attractive to practitioners for a number of reasons, such as its value in providing insight into the actions and intentions of users, and its ability to capture real-time responses from users during the testing process. Perhaps the main reason for its popularity among usability practitioners, however, is that it is fast and easy to implement (McDonald et al., 2012). The critical importance of time and cost in the IT industry often means that usability practitioners must conduct tests according to tight deadlines, and with limited resources at their disposal (Norgaard and Hornbæk, 2008). It follows, then, that the most popular testing method would be one that enables practitioners to carry out usability analyses and deliver their reports in a time- and cost-effective manner.

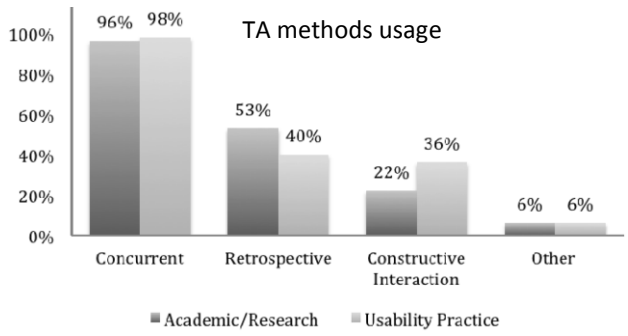


Figure 2.6: The usage of TA methods in research and practice (McDonald et al., 2012)

There are, however, several issues to be aware of which could have a negative impact on the quality of the data being collected, and these raise questions about the verbal reports generated under concurrent TA conditions.

The first of these issues concerns the completeness of the data collected. Ericsson and Simon (1998) acknowledge that although the concurrent data can provide sufficient evidence for the accurate sequence of thoughts that participants had whilst completing the task, the verbal reports are likely to be incomplete since participants are expected to give priority to task solving and may therefore fail to report some thoughts (Ericsson and Simon,

1998; Ericsson and Fox, 2011). This issue of completeness has been discussed at length in the psychological literature. A number of psychologists (Wilson, 1994; Schooler and Fiore, 1997; Wilson, 2004; Schooler, 2011) claim it is unlikely that TA protocols offer a complete representation of people's cognitive process. For example, they fail to capture information relating to unconsciousness, automatic processes, and those thoughts that are difficult or impossible to verbalise. Within the context of usability testing, research investigating the relationship between eye movements and TA protocols suggests that verbal reports may indeed be incomplete. Cooke (2010) found evidence from eye movement data to suggest that when participants were silent, they were still actively engaging in scanning and assessing different options on the screen for task solving. She concluded that it was unlikely that the CTA method could provide a full picture of users' interaction and their experience.

The second issue is simply that the process of concurrent verbalisation may feel uncomfortable or unnatural, as people do not commonly verbalise their thoughts constantly while working (Nielsen, 1993a).

The third issue concerns the extent to which the request to think aloud may interfere with and alter participants' thought processes. This may affect task performance, which in turn can affect the validity of the data obtained. This issue is often referred to as reactivity (Freeman, 2011). Within the usability community, the possibility of reactivity when using the CTA method has been discussed in a number of studies (e.g. Van den Haak et al., 2004; Hertzum et al., 2009;), although the specific term "reactivity" is not always used. In usability testing, reactivity poses a threat to the validity of data, as validity in this context is related to the extent to which the resultant data can be considered representative of real world use (Hertzum and Jacobsen, 2001; Blandford et aI., 2008). If participants demonstrate improved performance, this may result in the evaluators failing to identify problems, or to assign inappropriate severity ratings. If the participants' performance is impeded, it may introduce false problems (Zhao et al., 2012). Usability studies which have compared CTA with a silent condition alone or a silent condition followed by a retrospective thinking-aloud have had mixed results. For instance, Oostendorp and De Mul's (1999) found that the act of thinking aloud affected participants' task performance in the CTA condition when compared to participants in a silent condition. Other studies,

however, found no evidence of reactivity at all (e.g. Van den Haak et aI., 2004; Hertzum et aI., 2009). This limits the ability to draw firm conclusions about this issue.

Some researchers argue that reactivity does not result from the TA method *per se*, but from the fact that those studies reporting reactivity have deviated from Ericsson and Simon's classic framework (Meissner and Brigham, 2001; Ericsson, 2002; Fox et aI., 2011). Following an extensive review of the relationship between elicitation techniques and the validity of the resultant protocols, Ericsson and Simon (1993) published a framework on how researchers might elicit valid and reliable verbal data. In their framework, they recommend four procedural factors for TA data collection:

1. Minimal interactions between evaluator and participants. The evaluator should only issue TA reminders if participants fall silent, and the reminders must be short and non-directive, such as "keep talking", to safeguard against reactivity and evaluator-induced bias.
2. The provision of TA practice, which serves to increase participants' familiarity with the technique.
3. The use of general and neutral TA instruction, since instructions specifically requesting particular types of information may invite reactivity and yield inaccurate data.
4. Whenever possible, both concurrent and retrospective verbal protocols should be collected to enrich and enhance the accuracy of data.

2.6.2.2 Retrospective Think-Aloud Method

Retrospective think-aloud (RTA) is a method in which the users are asked to verbalise their thoughts after performing the tasks. The method has received less attention compared to the CTA (see Figure 2.6) (McDonald et al., 2012). Given the practical benefits of the concurrent method outlined in the previous section, and the fact that RTA increases the length of test sessions, why might practitioners consider the retrospective approach? The answer to this question is twofold. First, as mentioned earlier, questions have been raised about the validity of concurrent reports. Second, there are a number of benefits to using RTA protocol instead of CTA protocol. One such benefit is a possible decrease in reactivity: participants are fully enabled to execute a task in their own manner and at their own pace, and are therefore less likely to perform better or worse than usual. A second

advantage would be that since RTA participants verbalise their thoughts after completing tasks, they may have more opportunities to reflect on their experience of working with the test object (Cotton and Gresty 2006). Finally, with regard to usability testing which is carried out across cultures involving multiple languages, RTA may be an appealing alternative to CTA, since it is probably less difficult for participants to verbalise their thoughts in a foreign language after their task performance than to do so while working (van den Haak et aI., 2004).

RTA methods do have some drawbacks. One of these relates to the method's reliance on human memory, which is fallible: with the best of intentions, participants might forget specific things that occurred during a task. Ericsson and Simon (1993) state that some information may be lost in the case of retrospective research, which was confirmed by Peute et al. (2010). In an effort to tackle this issue, usability researchers and practitioners nowadays tend generally (but not always) to offer participants a visual stimulus (e.g. in the form of a video recording of their performance) to help them recollect their thoughts and experiences from the test session (van den Haak et aI., 2004).

Another drawback concerns the fact that participants may produce biased accounts of the thoughts they had while performing the tasks, i.e. participants may attempt to explain or justify their behaviour with logical, plausible reasons that may not necessarily reflect the truth (Cotton and Gresty 2006). However, there is evidence to suggest that this issue is extremely rare. Guan et al. (2006) examined the congruence of retrospective reports with participants' eye movements collected during the completion of four tasks in a usability test. They found the verbalisations to be an accurate reflection of what participants did during the task performance phase, with only 3% of verbal reports being inaccurate.

2.6.2.2 The Hybrid Method

In their influential work on protocol analysis, Ericsson and Simon (1984, 1993) suggest that both concurrent and retrospective protocols should be combined. They argue that this combination of both concurrent and retrospective verbal protocols, referred to as the *Hybrid* (HB) method in this thesis, can enrich the utility and enhance the validity of the verbal data collected. The issue of incompleteness associated with the CTA method could be addressed by the collection of retrospective data. In addition, gathering both types of

data can help assess the accuracy of the verbal data, as Ericsson and Simon (1993) argue that the information contained in concurrent and retrospective protocols should essentially be the same.

Surprisingly, at the present time, only a few usability studies have examined the combined use of concurrent and retrospective reporting in the same test. This may be because, as mentioned in section 1.2.1, the concurrent and retrospective think-aloud methods have evolved as separate approaches within usability testing, and are therefore more often compared than combined. The use of Ericsson and Simon's HB method in usability testing was first investigated by Følstad and Hornbæk (2010). For each task, they followed a classic CTA phase with an interpretation phase. In this second phase, the evaluator led a discussion with the participants about the important usability issues that occurred during the testing session. The researchers then carried out a comparison of the problems sets yielded by the concurrent reporting session, and the interpretation session. The results suggested that the interpretation session enhanced the CTA data by adding new problems that were not detected in the CTA phase. Although the authors referred to their second phase as "retrospective probing", their description of this approach gives the impression that the interpretation phase more closely resembled a post-test interview than RTA.

A more recent study by McDonald et al. (2013) examined the utility of the HB method. They divided the test session into an interaction phase and an interpretation phase. In the first phase, the participants were asked to think aloud while solving tasks, and once all the tasks were completed, they were invited to an interpretation session. In this session, the participants were asked to report retrospectively on each of the tasks. The results suggested that this second phase, after the concurrent think-aloud task solving, generated additional insights into the reasons behind the difficulties encountered and decisions made during task performance. However, neither of the aforementioned studies compared the HB method to any of the one-phase methods (such as CTA or RTA) to truly determine the utility of the approach.

There are several TA methods in usability evaluation practice that go beyond the traditional protocols. The following section will elaborate on these methods.

2.6.3 Relaxed Think-Aloud Methods

Relaxed TA methods refer to a range of variations on the classic TA method which have been created by usability practitioners or suggested by researchers. In these adjusted approaches, the evaluator plays a more active role than in the traditional method.

2.6.3.1 Active Intervention Method

Field studies have revealed a significant gap between the theory and practice of TA methods. In practice, a test evaluator will often actively encourage participants to talk about their intentions, thought processes, understanding, and mental model. This is accomplished through prompts and interventions that are much more intrusive. The survey conducted by McDonald et al. (2012) indicated that the majority of usability practitioners often utilise a wide range of interventions during the CTA testing process. Shi's (2008) observation of six usability tests in five companies in Beijing also noted a similar phenomenon. Nørgaard and Hornbæk (2006) observed fourteen CTA testing processes in seven different organisations, and found that the evaluators often used a relaxed approach, referred to in this thesis as the *Active Intervention* (AI) method, when it comes to intervening. Boren and Ramey (2000) used a combination of observations and interviews in their influential study of two professional usability companies, and found that the practitioners in their study often asked probing questions to seek explanations or clarify comments, rather than limiting themselves to the "Please Think Aloud" reminder.

While these studies have exposed a divergence between Ericsson and Simon's (1993) advice and how the TA method is applied by usability practitioners, an analysis of research studies investigating the use of TA methods reveals a similar pattern of misuse. There seem to be many usability researchers who fail to comply with the guidelines proposed by Ericsson and Simon, especially those that call for minimal interaction with participants. In fact, even some well-known handbooks on usability testing (Rubin and Chisnell, 2008; Dumas and Redish, 1999) encourage test evaluators to seek explanations and additional details from participants, since this might help to gain more insight into the deficiencies of a particular test object. While many usability practitioners readily take the advice offered to them in handbooks, there is no empirical evidence supporting the usefulness of interventions in enhancing the utility of collected TA data, particularly in terms of the number and severity of usability problems identified. As mentioned in section 2.6.2.1,

Ericsson and Simon (1993) believe that interventions and questions from evaluators can affect participants' verbalisations and task performance and impact the validity of data. Indeed, exploratory studies in the context of usability testing, such as Hertzum et al. (2009), and Olmsted et al. (2010), discovered that these "probing" TA protocols measurably affect the behaviour, performance, and satisfaction of participants when compared with classic TA methods.

2.6.3.2 Speech Communication Method

The difference between traditional CTA as theorised (Ericsson and Simon, 1993) and the actual practice of usability professionals has caused some researchers to wonder whether another approach to TA protocol might be more effective. Boren and Ramey (2000) suggest that a TA protocol based on speech communication theory, referred to here as the *Speech Communication* (SC) method, may be better suited to usability research. Boren and Ramey (2000) state that for usability studies, the traditional TA protocol where the test evaluator remains silent outside of short assertive commands to "keep talking" might be more disruptive to the participant than previously acknowledged, because humans communicate within a speaker/listener relationship. They argue that their protocol reflects the way human beings naturally communicate, with a combination of statements offered by a speaker followed by feedback or acknowledgment from a listener. According to speech communication theory, during a conversation, it is essential for the listener to use verbalised sounds or phrases which affirm to the speaker that the listener is paying attention and is absorbed in the communication act. The speaker's role (participant) is to talk and to offer information while the listener's role (test evaluator) is to respond as much or as little as necessary. This two step information exchange establishes an interaction between speaker and listener (Boren and Ramey, 2000).

The development of the speech communication protocol was also motivated by a review of the differences in purpose between research into cognitive processes and research into usability testing. When TA protocols are employed in cognitive psychology research, the focus of attention is the participants' cognitive process. When TA protocols are employed in the context of a usability test, the focus of attention is not so much the test subject as the system with which the subject interacts. So, essentially, there are two interactions taking place: one between the subject and the system, and one between the experimenter/evaluator

and the subject. Boren and Ramey (2000) indicate that the primary focus of usability testing is not to develop solid models of human cognitive processes, but to identify deficiencies in a particular test object. This means that only certain aspects of participants' cognitive processes are of interest to usability testers.

Given that usability practitioners have very different reasons for conducting TA tests than cognitive psychologists, Boren and Ramey's (2000) speech communication protocol allows the evaluator more freedom to interact with participants with the aim of better facilitating product evaluation rather than investigating participants' thought processes. Boren and Ramey recommend the use of acknowledgement tokens such as *"Mm hmm,"* as they can provide the expected response from an active listener whilst remaining non-directive. Since these acknowledgment tokens carry almost no content, they require little cognitive processing in order to be received and comprehended. The tokens are natural continuers and do not infringe upon the flow of communication. The evaluator should not ask questions directly or start a conversation. If the participant does fall silent, Boren and Ramey suggest that a practitioner employ the token "*Mm hmm?"* despite there being nothing to be acknowledged. If the participant continues to remain silent, then a neutral, content-free probe such as "*And now…?"* may be a more obvious prompt to maintain conversation (Boren and Ramey, 2000).

Although the SC protocol was designed with usability evaluation in mind, there is no definitive evidence regarding its real contribution, as no research has examined it in detail. To date, there has been only one study that compared the traditional TA protocol with the SC method (Olmsted-Hawala et al., 2010). More information about this study can be found in section 2.7, which discusses prior comparative studies of TA methods.

2.6.4 Co-Participation Method

Another increasingly common protocol in the context of usability testing is the *Co-Participation* (CP) method (see Figure 2.6). This protocol is also known as the *constructive interaction* or *team TA* method (e.g. Dumas and Redish, 1999), and involves two participants working together to explore the test object and perform tasks. The paired participants are asked to engage in verbalizing as they interact with the system and one another. O'Malley et al. (1984) introduced this method into the study of human-computer

interaction in the mid-1980s. The CP method is considered an effective way of making TA test participants feel more natural test participants (Van den Haak et al., 2004, Nielsen, 1993a). Nielsen (1993a) further states that the CP method is especially suited to usability evaluations involving children as it better facilitates children's verbalisation than does the classical TA protocol. However, using two people for each test increases the cost of testing and the difficulty of finding a sufficient number of test participants (Als et al., 2005).

2.7 Prior Studies Comparing Think-Aloud Methods

With such a proliferation of different strategies and methods for eliciting participant verbalisations during TA usability tests, there is a clear need for comparative research into the effects, utility, and validity of the different methods. Consequently, a number of empirical studies have been conducted comparing the impact of various methods—classic concurrent TA, retrospective TA, relaxed TA, and co-participation—on test outcomes. This section presents an overview of these comparative studies.

Comparison of Classic Think-Aloud Methods

Ohnemus and Biers (1993) were the first to conduct a comparative study of the classic TA methods. They compared the test participants' performance and subjective ratings in three test conditions: CTA, RTA with reports completed right after the test, and RTA with reports completed on the following day. The results found no significant difference between the groups in terms of either task performance or subjective ratings of the system. However, this study was limited as it did not take into account the number and quality of problems detected which is a key aspect of usability testing.

Van den Haak et al. (2004) conducted a similar study 11 years later, comparing CTA, RTA (with reporting immediately after the test tasks), and the CP method. The results showed no significant difference in the total number of problems found, but the problems were detected differently: the retrospective condition revealed more problems through verbalisation, whereas the concurrent condition revealed more problems through observation. Even so, the study found no significant difference in the severity of problems detected, in the participants' overall task performance, or in their experiences with the TA test.

Another study by Peute et al. (2010) compared the performance of the CTA and RTA, and showed that the CTA method performed significantly better than the RTA in detecting usability problems. In addition, CTA was more thorough in detecting usability problems of a moderate and severe nature. That said, CTA was found to prolong the task processing time.

Comparison of Classic and Relaxed Think-Aloud Methods

There have been three comparative studies that have measured the validity of the relaxed protocols against that of the traditional CTA protocol. A study by Hertzum et al. (2009) compared the traditional and the AI protocols to a silent condition. It was found that the CTA approach had very little effect on task performance, whereas the AI method seemed to alter the participants' behaviour, causing them to browse and navigate more within and between the web pages. The results confirmed that classic TA testing yields valid data about the use of the evaluated systems provided the interaction between participant and test evaluator is kept to a minimum. AI, on the other hand, may not be a valid method for gathering data about users' performance, as it may be associated with increased reactivity.

A study by Zhao and McDonald (2010) compared the CTA method with AI method. The results showed that most of the test participants (17 out of 20) preferred the more interactive TA approach, although the increased number of interventions also distracted some of the users, leading to poorer performance. Finally, Olmsted-Hawala et al. (2010) compared three different TA methods: CTA, SC, and AI and used a silent condition as a control. The study was a between-subject study with 20 participants and 4 evaluators that each conducted the test without knowledge of the true goals of the study. Three outcomes were measured: accuracy (considered in terms of success or failure with the tasks), efficiency (considered in terms of task completion time), and satisfaction (measured using the subjective satisfaction score about the website used). The results showed that the levels of accuracy were significantly higher in the AI condition, where 60% of the tasks were completed accurately compared to the 30-40% observed in the other conditions. The AI protocol also produced higher satisfaction scores, as participants gave more positive scores in this condition compared to the others. In terms of efficiency, no significant differences were found between the test conditions, even when compared to the silent condition. The

researchers concluded that usability practitioners should use either the traditional or the SC method, because the AI protocol created reactivity.

Comparison of Co-participation and Single-participant Methods

Adebesin et al. (2009) compared the CP protocol with the CTA and analysed the effect of the CP method on task performance. They found no significant differences between the methods. Similar results were found by Als et al. (2005) who also studied the CP and the CTA, and they found that the CP method costs less than the CTA method in terms of the total time expended by the evaluator to conduct testing sessions and analyse results. They also found that the paired test participants detected significantly higher number of usability problems than did the single test participants. In contrast, Van den Haak et al. (2004) found no significant differences between the paired test participants and the single test participants in the number of problems detected or in the task performance measures, but the CP was rated more positively by its users.

Assessment of Comparisons

Assessments and comparisons of usability evaluation methods in general (including TA methods) have been subjected to heavy criticism (Hornbæk, 2010). Therefore, even though the studies conducted on assessing TA methods in usability testing have improved the understanding regarding the validity and utility of the methods, several gaps can be identified in the literature.

First, it is evident that there is a need for a thorough and holistic assessment of the methods. TA protocols have been evaluated based on a range of criteria, including usability problem identification (Peute et al., 2010), task performance metrics (Olmsted-Hawala et al., 2010; Van den Haak et al., 2004), participants' testing experiences (Zhao and McDonald, 2010), the cost of employing methods (Als et al, 2005), and the number of test participants needed to find a sufficient number of usability problems (Nielsen, 2000) (see Table 2.1). However, no existing research unifies all of these criteria into a single study. The failure of previous studies to combine evaluation criteria has resulted in conflicting findings and an incomplete understanding. This research argues that a holistic assessment is essential to the establishment of a systematic, coherent body of knowledge regarding the contribution of TA methods to usability testing.

Table 2.1: Overview of the comparative studies on think-aloud methods

Study	**TA methods**	**Points of Comparisons**							
		Task Performance	**Participant's Experience**		**Usability Problems**		**Cost of Methods**		**Sample Size Needed**
			TA Test	**Website**	**Quantity**	**Quality**	**Temporal**	**Financial**	
Ohnemus and Biers (1993)	CTA vs. RTA	√	×	√	×	×	×	×	×
Van den Haak et al. (2004)	CTA vs. RTA vs. CP	√	√	×	√	√	×	×	×
Peute et al. (2010)	CTA vs. RTA	√	×	×	√	√	×	×	×
Hertzum et al. (2009)	CTA vs. AI	√	√	×	×	×	×	×	×
Zhao and McDonald (2010)	CTA vs. AI	×	√	×	×	×	×	×	×
Olmsted-Hawala et al. (2010)	CTA vs. AI vs. SC	√	×	√	×	×	×	×	×
Als et al. (2005)	CTA vs. CP	√	×	×	√	√	√	×	×
Adebesin et al. (2009)	CTA vs. CP	√	×	×	×	×	×	×	×

Second, although the main purpose of usability evaluations is to uncover as many problems as possible, the author has only found two empirical assessments of the usability problems identified via the different TA protocols. This limited focus on problem identification supports the general critique that usability research is "in crisis" and has little relevance to practice (Woolrych et al., 2011; Wixon, 2003). Furthermore, a great number of usability evaluation studies in general have only considered the number of problems detected by a certain method (Hornbæk, 2010). Researchers argue that counting problems does not always benefit usability research, as it ignores the difference between the seriousness and types of problems, and their value for optimization (Hornbæk, 2010; Furniss *et aI.,* 2007; Wixon, 2003). Hornbæk (2010) also observes that previous studies have tended to focus on the individual problem level (problems detected per participant) to the exclusion of the final problem sets (problems detected per method) (e.g. Als et al., 2005), meaning that there is no means to have a full picture.

Third, despite the significance that the evaluator effect can have on the validity of the data, the majority of studies do not consider or discuss this factor (Hornbæk, 2010; Hornbæk and Frøkjær, 2008). Section 3.10 in the following chapter details the factors that were taken into account in this thesis in order to minimise the evaluator effect.

Fourth, in a similar vein to the discrepancy between TA theory and practice, an examination of usability studies utilising classic TA methods found procedural inconsistencies in the administration of TA protocols (Lewis, 2014). TA research in usability testing does not often conform to its most cited theoretical basis, the work of Ericsson and Simon (1980). For example, while some studies used a general instruction, asking participants to say everything aloud (e.g. Van den Haak et al., 2004), others used explicit instructions to request explanations (Peute et al., 2010), and other studies failed to report the instruction used (Ohnemus and Biers, 1993). In some studies, participants had been able to practice TA (Hertzum et aI., 2009; Olmsted-Hawala et aI., 2010), while others did not report whether or not this was the case (Van den Haak et aI., 2004; Peute et al.,2010). In general, the level of information provided about the application of the methods was often poor (Makri et al., 2011). Omitted details and incomplete reporting made it impossible to ascertain whether or not the research engaged in certain activities. Additionally, no existing studies have compared the three classic TA methods—CTA,

RTA, and HB—described by Ericsson and Simon (1993), by testing the same interface using the same experiment design and set of metrics.

Fifth, as the review of available literature in this section has shown, the number of comparative studies on the utility of different TA methods in website usability testing is still limited. All in all, there are considerably more uncertainties regarding the value and the optimal design of TA usability testing than is suggested in the numerous textbooks available. Many aspects of TA usability testing deserve serious and systematic research attention.

2.8 Summary

This chapter has set out the background and context for the research presented in the thesis. A review of the relevant literature has indicated that, while TA methods have been widely applied in usability evaluation, they are not fully understood. At the present time, there is no consensus as to the utility and validity of these methods, and a cohesive body of knowledge regarding their application has yet to be established.

3

RESEARCH METHODOLOGY

3.1 Overview

Research in the field of human-computer interaction (HCI) requires a methodology that will provide in-depth understanding and knowledge (Lazar et al., 2010). Creswell and Clark (2011) define methodology as the overall process or model applied by the researcher to conduct a study and fulfil pre-defined research objectives. Research methodology can therefore be regarded as an umbrella term for an overall blueprint for a study and the various components of that blueprint. In order to choose the most appropriate research methodology for a study and to "safeguard against making elementary errors" (Denscombe, 2007, p.1), researchers must examine available research methods, techniques, and designs.

The purpose of this research is to investigate the use of different think-aloud (TA) methods in website usability testing. Following on from the introduction to the research and the literature review in the previous chapters, this chapter seeks to justify the choice of research methodology for the 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, the evaluation objects, and the strategies used to analyse the data. Lastly, it concludes by considering the validity, reliability and ethical considerations of the research.

3.2 Research Paradigm

The word research is composed of two syllables, *"re*" and *"search*". The dictionary defines the former as a prefix meaning again, anew, or over again, and the latter as a verb meaning to examine closely and carefully to test, try or probe (Dawson, 2002). Together they form a noun describing a systematic and scientific "inquiry or investigation into a specific problem, undertaken with the purpose of finding answers or solutions" (Sekaran 1992, p. 21). All research relies on some underlying beliefs regarding what constitutes genuine investigation and which research methods and techniques are appropriate for carrying out such investigation. This "basic set of beliefs that guides actions" (Guba and Lincoln, 1994, p. 17) is referred to in the research community as a research philosophy or paradigm. Researchers should therefore be explicit regarding the philosophical assumptions underlying their research (Dawson, 2002).

Two major philosophical standpoints have been identified in the western tradition of scientific research: positivism (sometimes called scientific) and interpretivism (otherwise known as anti-positivist). Both standpoints are rooted in Classical Greek tradition, with Plato and Aristotle (positivists) on the one hand, and the Sophists (anti-positivists) on the other (Hirschheim, 1985). Each has been used with success in different domains. While positivism and interpretivism do share some similarities in terms of the research process, they make distinctly different assumptions concerning the acquisition of knowledge (Kumar, 2005).

Positivism argues that "the study of human behaviour should be conducted in the same way as studies conducted in natural sciences" (Kumar, 2005, p.12), which assume that reality is stable and can be observed and described objectively. This academic tradition places a "considerable trust in numbers that represent opinions or concepts" (Amaratunga et al., 2002, p. 19). On the other hand, the interpretivist paradigm is based on the belief that a strategy is needed to differentiate between people and objects in the natural sciences, as reality depends on people's subjective understanding and, therefore, can differ from one individual to another. This paradigm concentrates on the collection of non-numerical data – such as people's beliefs, understanding and attitudes to present a detailed description of the issue under study (Amaratunga et al., 2002).

Even though many scholars emphasise the importance of specifying a paradigmatic standpoint that is either positivist or interpretivist, there are circumstances wherein both paradigms can be combined (Gable, 1994; Lee, 1991). Indeed, some authors have called for a combination of positivism and interpretivism for the study of social phenomena in order to improve the quality of research (e.g. Rudy, 1985; Kaplan and Duchon, 1988).

This thesis takes a pragmatist view: namely, that the philosophical perspective adopted should be suited to the research aims and questions set out in Chapter 1. Since this research aims to examine the effect of TA methods in an objective and generalisable manner, quantitative data such as time spent on tasks by participants assigned to TA conditions must be collected. However, as this thesis also intends to capture TA verbalisations and to question participants about their experiences in order to arrive at a better understanding of the issues under study, qualitative measurements are also necessary. Accordingly, the

present study adopts a mixture of quantitative and qualitative techniques for data collection. The combination of quantitative and qualitative data collection is typically known as "mix-mode research" or "triangulation" and is likely to generate a broader picture of the phenomenon at hand, enable the validation of research findings, and remedy the limitations inherent in a single data collection technique (Creswell, 2009) (further details regarding data collection are set out in section 3.6). Bryman (1998) argues that once a research philosophy has been set out, it needs to be associated with actual works by selecting the most suitable method for the research. Accordingly, the following section addresses the method selected for the current research.

3.3 Research Method

It seems that the differences between the terms "methods" and "approaches" are philosophical, in many cases they are used interchangeably. It is, however, important to explore the differences between the methods and techniques by defining these two terms. 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 research, a number of research methods have been suggested. Lazar et al.'s (2010) taxonomy for HCI research methods consists of case studies, diaries, surveys, focus groups, ethnography, and experiments. The key features of these methods are set out below:

- Case study: obtaining in-depth data regarding a specific instance within a specific real-life context in order to arrive at observations regarding its behaviour and operation.
- Diary: participants are required to record events that they engage in throughout a period of time.
- Survey: groups of participants are questioned about their attitudes, perceptions, beliefs and behaviour regarding the research topic in order to obtain a snapshot of practices, situations or views at a particular point in time.
- Focus group study: a small group of participants are questioned about their attitudes and the reasoning behind those attitudes towards the research topic.
- Ethnographic study: deep immersion and participation in a specific research context to develop an understanding that could not otherwise be developed.

- Experimental method: manipulating one or more variables while attempting to measure others, in order to examine the effect of one or a set of independent variables on another dependent variable and the relationships between them.

The appropriateness of each of these methods for a given study depends on several factors: the philosophical underpinnings of the research, the purpose of the research, the advantages and drawbacks of the given method for that purpose, the time and resources available, and the researcher's experiences. The first five of the above methods are typically categorised as "descriptive methods", which seek to gather information on the characteristics of the research subject without manipulating any settings or variables. In contrast, experimental methods involve effecting changes upon one or more variables to assess their causal impact on any other variables related to the research topic (Lazar et al., 2010). O'Rourke and Hatcher (2008) stress that methods that are essentially non-experimental in nature provide little evidence regarding "cause-and-effect relationships", negating the possibility of drawing strong inferences from their findings. In light of this argument and the usefulness of experimental research in enabling the identification of causal relationships, the experimental method was deemed the most suitable for the present research.

In HCI, the experimental method originated from behavioural research and is largely rooted in the field of psychology. It currently plays a key role in HCI research, having led to many groundbreaking findings in the field (Lazar et al., 2010). However, many researchers have criticised this method in relation to issues of validity and reliability (these issues are discussed in more detail in section 3.12). An experimental study normally starts with a research question or a testable research hypothesis, which is "a precise problem statement that can be directly tested through an empirical investigation" (Lazar et al., 2010, p.12). Other components of experimental research include conditions and units. Conditions, also known as treatments, refer to the different techniques, factors, or procedures being compared, while units are the objects to which experimental conditions are applied. In HCI and usability research, units are normally human participants selected based on specific characteristics such as gender, age or computing experience (Lazar et al., 2010).

3.4 Research Design

A detailed research design is a pre-requisite for the success of any research project. Yin (1984, p.19) defines a research design as: "an action plan for getting from *here* to *there*, where *here* may be defined as the initial set of questions to be answered, and *there* as some set of conclusions (answers) about these questions". In other words, a research design sets out a systematic procedure for achieving the pre-defined goals of a study within a specified timeframe.

Figure 3.1 breaks down the research design of the current study into its constituent steps and phases from formalisation to conclusions. This research consists of three phases: research design; data collection and analysis; and discussion and conclusions. Each phase is highlighted in a different colour and is mapped to a set of research objectives in Figure 3.1. Ideally, the research will progress in the manner indicated by the small dark arrows in Figure 3.1; that is, each phase of the research will start only after the previous one is completed, meaning that the activities in each phase can be iterated to the researcher's satisfaction. However, if new findings emerge, the researcher may need to revisit previous phases; for example, the researcher may revisit the literature to compare the findings of this study to those of other researchers. The dashed arrows denote the feedback process and the possible backtracking process.

As illustrated in Figure 3.1, the starting point of the research process is a thorough and systematic review of usability testing literature, which provides a foundation for developing an understanding of the research area under investigation. From the literature review, several issues which require more focused attention are identified. This leads to a specific research area and ultimately, a research need. The recent literature has raised a number of issues concerning TA methods within the context of usability testing that merit further research (see section 1.3). As a result, the researcher was able to identify a specific problem to be investigated and the aims to be achieved, and to formulate a set of research questions. After conducting further reading of the literature, the researcher was then able to specify the most suitable research paradigm (mixed mode) and method (experimental) to answer the research questions, as discussed earlier.

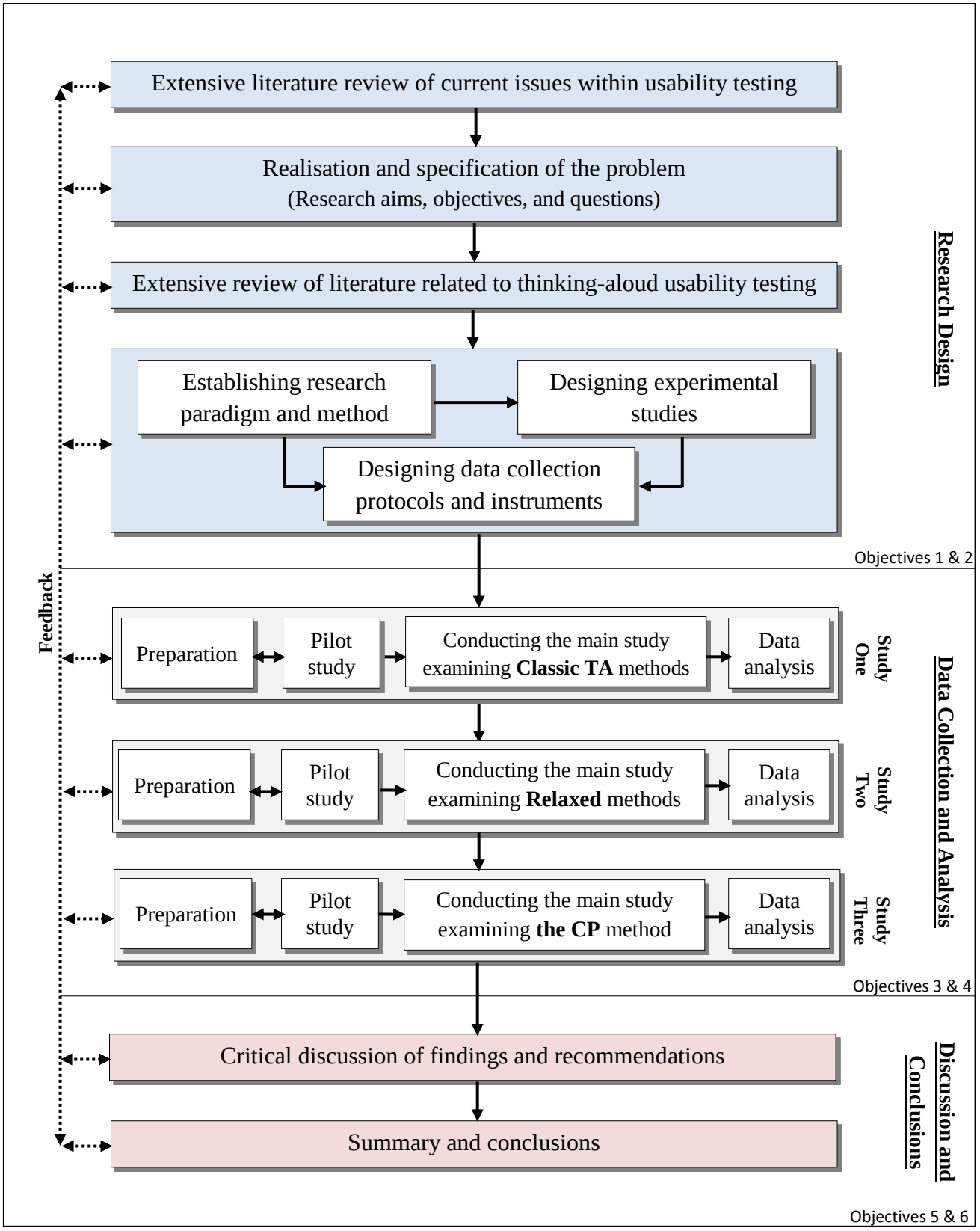


Figure 3.1: Research design and components

Other key tasks in the first phase of the research process are the identification of research variables, measurements, data collection techniques, and the design of the experimental studies and instruments. The following sections discuss these and other phases of the research design more thoroughly, beginning with a discussion of experiment design in the next section.

3.5 Experiment Design

There is a great difference between usability test design and the design of an experimental usability testing study. Usability testing aims to find flaws in a specific interface, whereas experimental studies address the effect of certain factors on the outcome of usability testing. The latter has at least two treatments and usually requires many participants to obtain meaningful data, and its results must be validated and reported to the scientific community. In order to obtain valid results, conducting the actual experiment must be preceded by a carefully planned process which includes the identification of variables, the determination of an experiment structure, and the selection of an experimental approach (Lazar et al., 2010).[1]

3.5.1 Variables

In any experimental study, it is essential to identify the independent and dependent variables between which a relationship may exist. Independent variables, also known as input or predictor variables, are manipulated by the researcher in order to answer specific research questions, and may affect other variables. Dependent variables, also called outcome or response variables, are those which are measured in the experiment, and are subject to the influence of independent variables (Sternberg et al., 2007).

In the present research, the independent variable under examination is the type of TA methods. The dependent variables are the following evaluating criteria of TA performance: 1) performance data from participants' tasks, 2) participants' subjective testing experience, 3) usability problem data, 4) cost of methods, 5) and the relationship between sample size and problem detected. These five themes have been identified by the researcher as being

[1] In the present research, the words "experiment", "test", "evaluation" and "study" are used interchangeably.

typical of themes used to investigate TA methods in usability research. Section 3.9 discusses these in further detail.

A well-designed experiment must also take into account factors other than independent variables that may affect dependent variables; such factors are known as confounding variables or third variables. Well-known examples of confounding variables in usability testing research include the environment in which the test is conducted, the test settings, and individual differences between participants. Failure to control confounding variables may lead to a false conclusion regarding the cause-and-effect relationship between independent and dependent variables (Lazar et al., 2010).

3.5.2 Experimental Structure

The second step of experimental design involves constructing the structure of the experiment based on the research questions that have been developed (Lazar et al., 2010). The main structure of an experiment can be determined by answering the two questions below:

- How many independent variables are investigated in the experiment?
- How many different values or groups are in each independent variable?

The answer to the first question decides whether a basic or factorial design should be adopted. If only one independent variable exists, a basic design must be adopted. However, if there are two or more independent variables, a factorial design must be used. The answer to the second question determines the number of conditions needed in the experiment (Lazar et al., 2010). As there is only one independent variable in each study (i.e., the type of TA methods), this research adopts a basic design. Since this independent variable has more than one value (i.e., different TA variations), this research involves multiple conditions. The first study examined the classic TA methods, the second study investigated the relaxed TA methods, and the third study explored the co-participation methods. The conditions are clearly outlined in each individual study chapter.

3.5.3 Experimental Approach

Following the structuring of the experiment and the setting of conditions, an experimental approach must be selected based on whether the same participants or different participants

will be used across conditions. The use of the same participants is known as a "within-group" approach, while use of different participants is known as a "between-group" approach. Selection of an experimental approach is a critical step in experiment design, as it directly affects the quality and statistical analysis of the data collected. This decision also allows the general scope of the study to be outlined and a reasonable estimate for the timeline and budget of the study to be created (Lazar et al., 2010). The strengths and weaknesses of these two experimental approaches are discussed below and set out in Table 3.1.

A *between-group approach*, also known as a between-subject approach, assigns different groups of participants to different experimental conditions. This approach is effective in preventing the "carry-over" effect which can result from learning (improving performance) or fatigue (decreasing performance), as participants are exposed only to the condition to which they are allocated (Lazar et al., 2010). However, when a between-group approach is adopted in usability evaluation studies, individual differences among participants such as demographic details may have a substantial impact on participants' performance (Olmsted-Hawala and Jennifer, 2012). In order to reduce the impact of individual differences, large and roughly equal numbers of participants with similar demographic features must be allocated to each condition. This leads to the second major disadvantage of this approach: large sample size (Howitt and Cramer, 2007).

In contrast, a *within-group approach***,** also referred to as a within-subject approach, requires each participant to experience multiple experimental conditions. This effectively isolates the impact of individual differences as all participants are exposed to all experimental conditions, and therefore does not require large sample size and is less resource intensive. The primary disadvantage of the within-group approach is the possible impact of "carry-over" effects. Since all participants undergo all experimental conditions, they are very likely to learn from the experience of the first condition; therefore, their performance under another condition may be improved in ways that do not accurately reflect the effect of that condition. The within-group approach also requires more time from participants, which may induce confounding factors such as mental and/or physical fatigue and frustration (Howitt and Cramer, 2007). Steps can be taken to reduce the impact of the 'carry-over' effect by allowing intervals of sufficient length between conditions, and

in some cases using counterbalancing techniques wherein participants are divided into groups and conditions are administered in a different order for each group, such as a "Latin Square" design (Lazar et al., 2010).

Table 3.1: Advantages and disadvantages of between-group design and within-group design (Howitt and Cramer, 2007)

	Between-group design	**Within-group design**
Advantages	Avoid learning effect Better control of confounding factors Cleaner	Small sample size Effective isolation of individual difference More powerful tests
Disadvantages	Impact of individual differences Harder to get significant results Large sample size	Hard to control learning effect Large impact of fatigue

Considering the advantages and disadvantages of both approaches, the between-group approach was chosen as the most appropriate experimental approach for the current research. The within-group approach was rejected because of the possible "carry-over" effects between the TA conditions of each study. For instance, participants could have provided more verbalisations than they would otherwise have provided due to increasing familiarity with the TA process, or could have become aware of the purpose of the study. Indeed, the majority of comparative TA studies favour the between-group approach (e.g. Van den Haak et al, 2004; Olmsted-Hawala et al., 2010; McDonald et al., 2013).

3.6 Overview of Data Collection

As mentioned earlier, the research questions required the collection of both qualitative and quantitative data. Hence, a triangulation of quantitative and qualitative data collection techniques was applied. The data collection involved two stages in each individual study: the first stage (the pre-study stage) collected data from participants through a pre-study (screening) questionnaire (Appendix C6) in order to recruit suitable candidates and control individual differences. The second stage (the during-study stage) involved three data collection techniques: observing participants' interactions with the system during testing, listening to participants' verbal comments (TA protocol), and collecting participants' answers to post-experiment questionnaires. These data collection stages are illustrated in Figure 3.2.

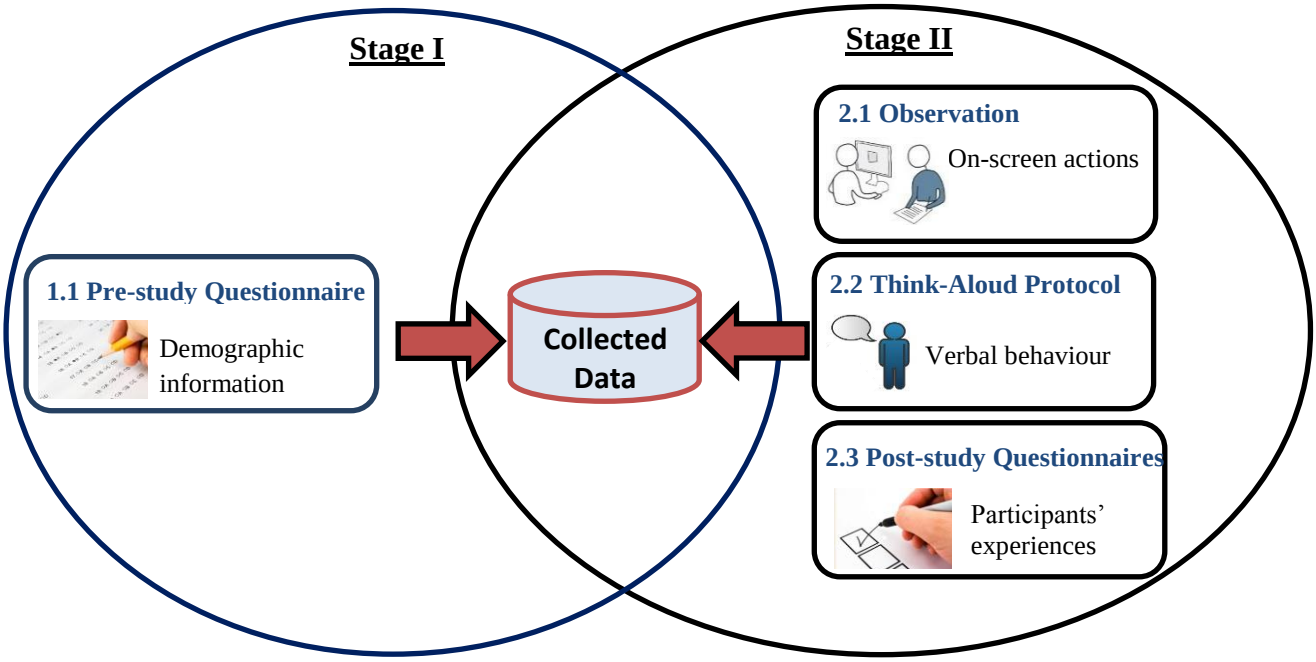


Figure 3.2: Data collection process

The following subsections elaborate further on each of the data collection techniques employed for the current research.

3.6.1 Observation

The observation technique involves gathering real time data on people' behaviour relating to a specific phenomena (Lazar et al., 2010). Broadly speaking, there are two primary types of observation techniques: covert and overt observations (Saunders et al., 2007). Covert observation occurs when the participant does not know that they are being observed. As the observer is hidden, participants are expected to act more naturally, though this may raise problematic ethical issues such as the lack of informed consent (Parker and Sara, 2014). Conversely, in overt observation the participant knows they are being monitored; this fact might affect their behaviour, in a phenomenon known as the 'Hawthorne effect' (McCambridge et al., 2014), particularly if they are very concerned about being observed. That said, Macefield (2007, p.9) argues that the "Hawthorne effect" is a "controversial idea that has highly questionable relatability" to the usability evaluation discipline, as "there are many significant differences between the studies carried out at Hawthorne works and typical usability studies". Due to the nature of the research problem, it was deemed necessary to use the overt observation technique. This technique enabled the researcher to

address questions from participants and deal with issues arising during the test session. Additionally, it was believed the researcher's presence would make participants feel less self-conscious about thinking aloud (Rubin and Chisnell, 2008). The data collected through overt observation mostly consists of participants' interaction with the systems tested and their quantitative task performance data (for more details on performance data, see section 3.9.1).

The researcher acted as the evaluator for all the thesis experiments.[2] Having only one person observing participants and evaluating their performance at the same time is generally acceptable, but can sometimes be problematic. However, if data analysis is based on video recordings, as in the case of this research, it is less problematic than when observation and analysis are carried out simultaneously (Jacobsen, 1999).

3.6.2 Thinking-Aloud Protocols

Participants' verbal reports will be derived from the TA protocols, which are the focus of this research. As mentioned on different occasions, such protocols enable participants to verbalize their thoughts with respect to their mental processes, impressions, and feeling about using a particular system. This in turn, helps the evaluators to understand how the participants undertake specific tasks, what kind of usability problems they encounter, and how they judge the quality of the system (Tullis and Albert, 2008).

3.6.3 Questionnaires

Questionnaires are one of the most commonly used data collection techniques across all research fields. In simple terms, a questionnaire is a range of questions designed to elicit answers from individuals to obtain information about a given topic. Questionnaires can be used for a range of purposes, such as describing populations, explaining behaviour, and collecting the opinions of participants regarding a particular phenomenon. When properly constructed and responsibly administered they can be a robust instrument yielding data with high validity (Lazar et al., 2010).

[2] From this point onwards in this thesis, the words "evaluator", "researcher", and "observer" are used interchangeably to refer to the author.

There are two general types of questionnaire: self-administered questionnaires and researcher-administered questionnaires. Self-administered questionnaires are completed by respondents in their own time with no researcher present, whereas researcher-administered questionnaires are completed by a researcher using the participants' responses or by the participants themselves but under the supervision of the researcher (Saunders, 2009). The questions themselves can be divided into closed-ended or open-ended questions. Closed-ended (structured) questions ask individuals to give a specific answer using few words or select an answer from a given set of choices; open-ended (unstructured) questions ask individuals to provide a response in the way with which they are most comfortable (Lazar et al., 2010).

In this research, questionnaires were used for two purposes. Firstly, as mentioned earlier, a self-administered screener questionnaire was sent to participants in advance of each study to gather demographic information. Secondly, researcher-administered questionnaires were employed at the end of each experiment to assess participants' experiences of the testing environment and their satisfaction with the tested website. The screener consisted of a mix of open and closed questions (attached in Appendix C6), the majority of which were closed questions. The post-study questionnaires made use of a five-point Likert scale (attached in Appendices B1 and B2). Section 3.9.2 provides further details on the design process and the content of the questionnaires, with particular attention to the post-study questionnaires.

3.6.4 Secondary Data

The three data collection techniques outlined in the above subsections, namely, observation, TA protocols, and questionnaire, served as the main sources of primary data for the present research. Primary data consists of first hand data collected expressly for a study by the researcher from original sources. The other form of data, secondary data, consists of data readily available in the public domain. Such data are normally inexpensive and can be obtained from many sources, including textbooks, academic journals, electronic sources, and newspapers (Krathwohl, 1997). In this thesis, secondary data is derived from the literature review and contributes to the design and implementation of the study. The researcher was able to examine numerous publications via hard copies in the University of East Anglia (UEA) library and by using an Athens account provided by the university.

Some top-level databases that may be of interest to HCI and usability researchers are shown in Table 3.2. A further list of journals and periodicals is provided in Sauro's (2013) *17 Periodicals for Usability Research*.

Table 3.2: Databases of potential interest to HCI and usability researchers

Database	Main Content
Journal of Usability Studies	Empirical findings, usability case studies, the practice and education of user experience
International Journal of Human-Computer Interaction	Cognitive, creative, social, health, and ergonomic aspects of interactive computing
Journal of Interacting with Computers	HCI and design theory; new research, interaction process and methodology; user interface, usability and UX design
Journal of Computers in Human Behaviour	HCI, the use of computers in psychology, the psychological impact of computer use on individuals, groups and society
CHI Conference Proceedings	Cognitive psychology, design, social science, human factors, artificial intelligence, graphics, visualization, multi-media design
INTERACT Conference Proceedings	Methods and tools for interface and interaction design, modelling, and evaluation, cross-cultural and social issues
HCII Conference Proceedings	HCI, human interface and the management of information

3.7 Test Objects

The test objects in this thesis are digital university libraries. Of the many different views in the literature on what constitutes a digital library, perhaps the most widely cited definition is that of Arms (2000, p. 2), which describes digital libraries as a "managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network". For universities, online libraries are an increasingly important channel to library resources and services targeting a broad group of students and other potential visitors, such as faculty and library staff. The popularity of such websites stems mainly from their reduction of spatial and temporal barriers by enabling users to search and browse their collections at any time from any location via the Internet. Users of these websites should be able to achieve their goals efficiently, which means that information should be easy to find, comprehensible, and supported by clear design. However, many users experience obstacles on these websites that hinder the efficiency of reaching their goals (Jeng, 2005). Furthermore, targeting university online libraries was also expected to facilitate the process of finding a truly representative sample

of users for the targeted sites, which in turn would facilitate the selection of research participants.

While empirical evidence on the effect of TA methods on the usability testing of websites has been limited (Olmsted-Hawala et al., 2010), this dearth of evidence is more visible with regard to academic library websites. Based on the above premises, the researcher therefore decided to focus on university library websites as test objects for the thesis experiments. The sections entitled "Test object" in Chapters 4, 5, and 6 provide more details about the specific website chosen for each study.

3.8 Choice of Setting

The setting of the thesis experiments required careful attention due to its profound importance for this research and, more generally, for any experimental study in the HCI field (Kjeldskov et al., 2004). One consideration with regard to setting is whether to carry out experiments in the lab or in the field. Conducting usability experiments in the field may allow researchers to discover unanticipated phenomena and study activities too complex to bring into the laboratory, but can also decrease researcher control over the study. Conversely, a laboratory setting increases researcher control over experiments; minimises the effect of external influences (e.g., environmental conditions; the speed of Internet connections); facilitates the process of data recording; and removes the need for researchers to travel to participants (Rubin and Chisnell, 2008); these advantages may partially explain why TA usability testing methods are more often applied in laboratory settings (Norman and Panizzi, 2006). However, Johnson (1998) criticises the use of laboratory experiments in HCI for their artificial settings. Considering the benefits and drawbacks of the field and the laboratory experiments, it was deemed more appropriate for the comparative studies to take place in a laboratory. In fact, TA usability testing is usually referred to as "laboratory usability testing" (Hartson et al., 2001, p. 374) which means that the form of the thinking aloud usability testing is regarded as the same as that of the lab experiment.

3.9 Measurements

While researchers have applied a number of measures to investigate the use of TA methods in usability testing, no previous research has taken a holistic approach to this issue, leading to a dearth of knowledge on the contribution of such methods to usability testing. This study takes a holistic approach to assess the TA methods in question in terms of both utility and validity. Utility refers to the usefulness of a method in assisting usability work, whereas the validity of a method refers to the degree to which the data collected conform to the real-world use of the system under study (Blandford et al., 2008). As mentioned in Section 3.5.1, five dependent variables are measured in this research: task performance data, participants' testing experiences, usability problem data, the cost of employing methods, and required sample size to find sufficient usability problems. These variables and their measures are discussed in more detail below.

3.9.1 Task Performance

Task performance measures are often used to assess reactivity associated with TA methods (Hertzum et aI., 2009; Olmsted-Hawala et aI., 2010). Participants' task performance measures collected in this research comprise task completion rate, time spent on tasks, and navigational behaviour.

Task completion rate is a widely used performance measure which quantifies the percentage of tasks completed correctly during testing (Tullis and Albert, 2008). The scheme used for categorising task completion, presented in Table 3.3, was constructed based on Tullis and Albert's (2008) coding scheme.

Table 3.3: Categorisation scheme for task completion (Tullis and Albert, 2008)

Category	Definition
Completed	Completed successfully
Failed	Participant gave up Participant performed the task incorrectly Participant believed that the task was complete even though it was not

Time-on-tasks, sometimes referred to as task completion time or simply task time, measures the time it takes a participant to perform a single task from start to completion and is usually expressed in seconds or minutes (Tullis and Albert, 2008). In the current

research, time-on-tasks was obtained for each individual task and for the completion of all tasks.

Navigational behaviour included the pages that each participant browsed and the number of mouse clicks that occurred during their browsing. Such data can offer greater insights into the influence of TA methods on user behaviour (Hertzum et al. 2009). These data were collected in Study Two and Study Three of this research.

3.9.2 Participants' Experiences

Two questionnaires are employed in this research to measure participants' subjective experiences: experience with the TA test questionnaire, and the System Usability Scale (SUS) questionnaire.

The experience with the TA test questionnaire aims to understand participants' experiences of the TA testing environment (Appendix B1). Measuring participants' testing experiences investigates the ecological validity of the TA variations under study. Ecological validity is concerned with the extent to which to a method is comfortable for participants to use. It is important for usability evaluators to ensure this type of validity, as test participants who feel stressed or uncomfortable about participating might encounter more problems than they should, or may fail to report usability problems that they would normally have noticed outside a test situation (Van den Haak et al., 2004).

The experience with the TA test questionnaire (Appendix B1) was based on previous research (Van den Haak et al., 2004). Four experts were asked to review the instrument: an English language professional and three scholars in TA usability testing. The TA testing referees were chosen on the basis of their willingness to evaluate the instrument, their ability to communicate the required information quickly, and their several years of experience in TA usability testing. Minor changes were then made to the questionnaire according to their suggestions. In addition, all the questionnaire items were piloted before their actual use in the first study to ensure that their wording would not introduce any potential biases and to assess the time needed for filling in the questionnaire. The questionnaire (Appendix B1) was structured and contained ten measurement items focusing on three elements of testing: 1) participants' views on how the method they used

affected their normal working procedure (in terms of speed and focus), 2) participants' opinions regarding the TA experience (e.g. the naturalness and ease of thinking aloud), and 3) how the presence of the evaluator affected their experiences. For each of these three elements, participants rated their experiences by indicating the extent to which they agreed or disagreed with a number of statements on a five-point scale, with a rating of 1 for "strongly disagree" and 5 for "strongly agree", as recommended by Lazar et al. (2010). This scale provides answers in the form of coded data that are comparable and can be readily manipulated. A sample statement is shown in Figure 3.3:

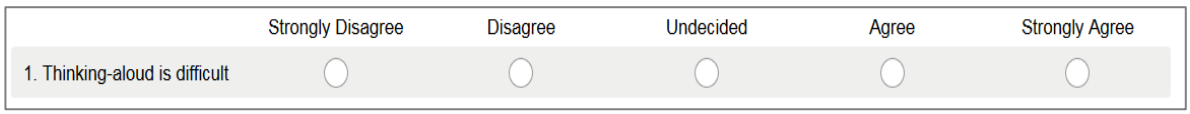


Figure 3.3: Sample statement from the participants' testing TA experience questionnaire

The SUS questionnaire developed by John Brooke in (1986), was used in this research to investigate the effects of the variations of TA protocols on participants' satisfaction with the tested websites. The questionnaire contains ten items with 5 response options (see Appendix B2). A sample statement is shown in Figure 3.4:

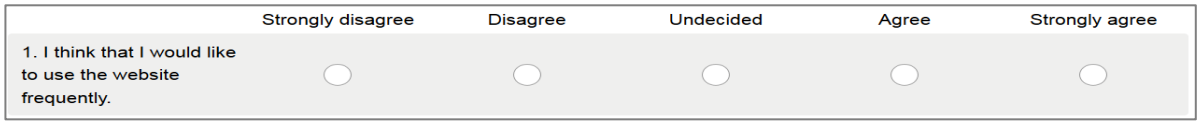


Figure 3.4: Sample statement from the SUS questionnaire

3.9.3 Usability Problems

Identifying usability problems is typically the primary purpose of usability testing (Hartson et al., 2001). Even though, there is no uniform definition of a usability problem, the current research project refers to the widely used definition of Lavery et al. (1997, p.7):

> "*an aspect of the system and/or a demand on the user which makes it unpleasant, inefficient, onerous or impossible for the user to achieve their goals in typical usage situations.*"

Problem counting: The most common way to measure the utility of usability evaluation methods (UEMs) is to count the number of problems they identify. This is frequently referred to as the *thoroughness* of a method (Hartson et al., 2001). To measure the thoroughness of the TA methods under study in the current research, the proportion of the

usability problems found by each TA method to the total number of usability problems found by all methods in each study was assessed. However, several researchers state that research on the utility of UEMs should not only focus on the number of problems produced, but also on the qualitative differences of these problems (Hornbaek, 2010; Blandford et al., 2008; Wixon, 2003). In line with these recommendations, this research assesses the quality of problems in terms of their sources, severity levels, types, and uniqueness:

Problem source: This term refers to the evidence used to find usability problems. Usability problems are easiest to detect from verbal data, as this requires less interpretation on the evaluator's part. Some problems can be detected based solely on the evaluator's observations; however, such detection relies significantly on the evaluator's judgement, increasing the likelihood that problems will be missed by the evaluator. Other usability problems can be detected from a combination of verbalised evidence and observed behaviour (Van den Haak et al. 2004). This research's investigation of problem sources seeks to determine how different TA methods can affect an evaluator's ability to identify and understand problems.

Problem severity: Molich and Dumas (2008) argue that it is more useful to locate severe problems than to find "all" problems, as problems with higher impact are more likely to be fixed by designers than those with lower impact. Thus, UEMs that uncover a high number of severe problems are more valuable than those uncover a high number of minor problems (Lindgaard and Chattratichart, 2007). As Hertzum (2006) notes, evaluators' assessments of problem severity may vary greatly and may not always be reliable. A common way to estimate the usability problem severity by the experts' judgements which can be done by asking usability specialists to rate the severity of each problem. Some researchers, however, hold that objectivity can be increased by ensuring that severity assessments are derived from user data rather than the evaluator's personal judgement (Hertzum, 2006; Lewis, 2006a), while others advise that problem severity should be assessed according to how participants' performance is affected (Nielsen, 1993; Dumas and Redish, 1999). In accordance with the above, the present research breaks down severity levels according to participants' task performance and based on the popularly used four level severity ratings (Dumas and Redish, 1999, Zhao et al., 2012), as outlined in table 3.4.

Such an analysis was derived from user data, and would therefore be less subject to the evaluator's personal understanding of the problems.

Table 3.4: Coding scheme for problem severity levels

	Problem Severity level	Definition
1	**Critical**	The usability problem prevented the completion of a task
2	**Major**	The usability problem caused significant delay or frustration
3	**Minor**	The usability problem had minor effect on usability, several seconds of delay and slight frustration
4	**Enhancement**	Participants made suggestions or indicated a preference, but the issue did not cause impact on performance

Problem type: Hornbaek (2010) and Blandford et al. (2008) suggest that while there is no single method that can effectively detect all usability problems, different methods can be more suited to detecting certain types of problems (e.g., navigation, layout, content). In this regard, Hartson et al. (2001, p. 110) state that "classification of usability problems by type is not only valuable within the usability development process, but is also necessary for characterizing the strengths and weakness of usability evaluation methods within usability evaluation methods comparison studies". Therefore, examining problem types can aid in revealing whether TA variations differ in their ability to detect different types of problems.

Unique and shared problems: Apart from the number, source, severity level, and type of problems, it is also important to analyse the uniqueness of the problems discovered. According to Law and Hvanneberg (2004), unique problems are those that are found only by one of the groups involved in testing, while shared usability problems are those detected by multiple groups. Addressing the uniqueness of problems discovered can help shed light on the differences between the problems discovered by participants in different TA conditions, and in turn provides further understanding of the ways in which they interact with the systems being tested.

3.9.4 Cost of Employing Think-Aloud Methods

An array of earlier studies, which conducted comparisons between TA testing methods and other evaluation methods, compared the cost of employing those methods (e.g., Martin et al., 2014; Hasan, 2009; Andreasen et al., 2007; Law and Hvannberg 2002; Molich and

Dumas, 2008). However, there is a lack of comparative study examining the cost of employing different variations of TA study. The cost of employing the TA methods is measured in the current research by recording the time spent conducting actual testing and analysing the results for each TA method. *Testing time*, recorded via an observation sheet (Appendix C18, Appendix D8), refers to the time taken to carry out the entire testing sessions, including the instruction of participants, data collection, and solving problems that may arise during test sessions. *Analysis time*, collected via web-based free time tracking software called "Toggle" [3](Version 2013), refers to the time taken to extract the usability problems from each method's testing data. The most efficient TA method can be determined by comparing the time and effort spent by the evaluator during each stage of a study. The less time and effort spent conducting testing and analysing results, the more efficient the TA method become.

The collected data above were also utilised for a comprehensive evaluation of the financial costs of the testing methods. According to Martin et al. (2014), usability professionals charge £800.00 per 7.5-hour day for usability consultation. This means that the hourly fee for usability consultation is approximately £107. This figure can be incorporated into the collected data to calculate the total costs of applying each TA testing method in a business environment. By comparing the financial cost of each method against the amount of usability problems found by each method, the cost per problem can also be deduced and compared (Als et al., 2005).

3.9.5 Relationship between Sample Size and Problems Detected

The last research question in this study (see section 1.5) focuses on the relationship between sample size and the number of problems detected, and in particular seeks to investigate whether sample sizes work differently for the TA methods under investigation. As mentioned in section 2.5.2, the issue of optimal sample sizes for usability testing has long been a subject of heated debate in the literature. Nielsen (2000) has controversially suggested that five participants are sufficient to uncover 85% of usability problems. Thereafter, Virzi's (1992) law of diminishing returns seems to apply, as fewer and fewer new problems are identified by involving additional participants (Virzi, 1992). Many

[3] https://toggl.com/

articles, however, with titles such as *Why Five Users Aren't Enough* (Woolrych and Cockton, 2001) and *Eight is Not Enough* (Perfetti and Landesman, 2002) critique the five-participant assumption by expressing concern regarding the impact of usability problems that may be missed when a group of only five participants is involved. Most of these articles have focused on the CTA method and no research has yet led to conclusive results. Accordingly, this research aims to explore in depth the effects of sample size on the number of usability problems detected.

Figure 3.5 below presents a visualisation of the dependent variables and their associated measures in what the researcher refers to as an "evaluation tree".

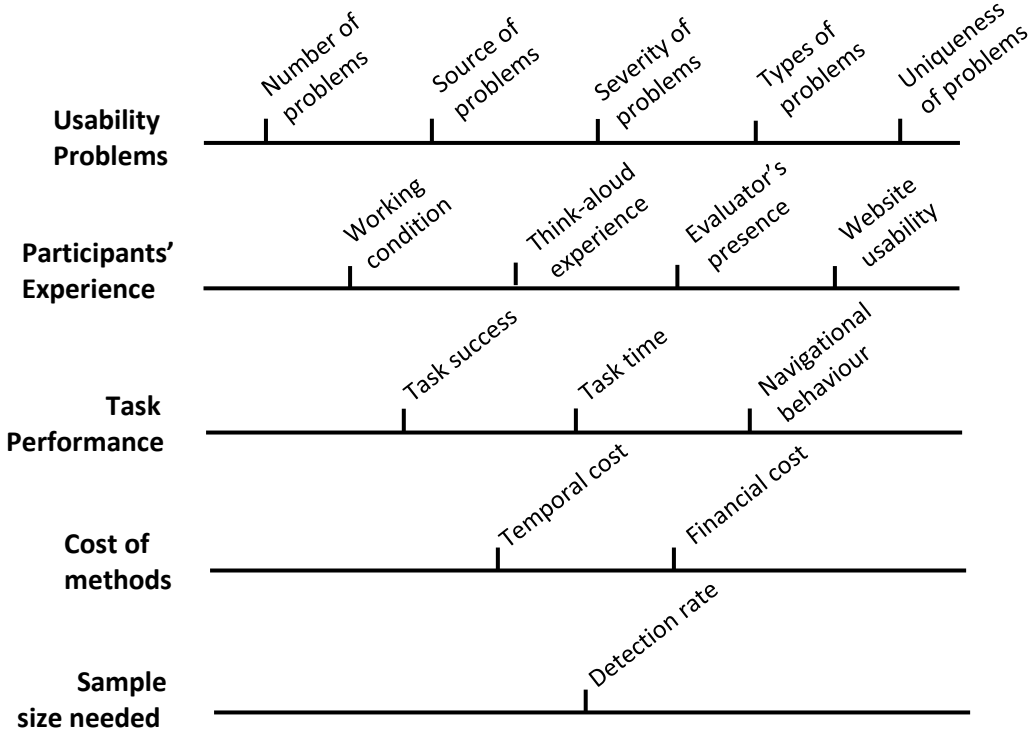


Figure 3.5: Visualisation of the evaluation criteria and measures of TA performance (evaluation tree)

3.10 Usability Problem Extraction

To date, there are no standard guidelines in existence for how usability problems should be extracted (Hornbaek, 2010). The literature on usability testing has paid little attention to this process in favour of examining the preparation phase and the conducting of testing sessions. Discussions of the extraction process tend to criticise it for its unreliability due

to the evaluator effect, which refers to the phenomenon wherein different evaluators when using the same evaluation technique to evaluate the same user interface identify different numbers of usability problems (Hertzum and Jacobsen, 2001). In the current research, a number of measures were considered during the problem extraction process based on recommendations from the literature in order to reduce the evaluator effect and to increase the reliability and validity of data.

Problem indicators: Research shows that the use of vague and non-uniform problem indicators in the problem extraction process can maximise the evaluator effect (Hertzum and Jacobsen, 2001; Hornbaek, 2010). It is therefore advisable for problem indicators to be clear and explicit. The critieria used to identify problems in usability studies have ranged in scope from short lists of less than 10 indicators (Jacobsen et al., 1998), to detailed checklists such as the Detailed Video Analysis (DEVAN) checklist (Vermeeren et al., 2002). This variability can lead to significant discrepancies between the numbers of problems discovered during test sessions. Hertzum and Jacobsen (2001) state that the use of vague, general criteria may be insufficient to guide problem extraction, causing evaluators to overlook certain types of problems. It is important to note, however, that the specific goals of a test, and the particular tasks being performed, may call for very different criteria to other, superficially similar tests. With this in mind, it is almost impossible to have a universally-applicable checklist of problem indicators. The development of a checklist should therefore be an iterative process, in which the criteria are continuously revised according to the needs of the practitioner. In response to these findings, this study applies a set of clear and explicit criteria to the process of problem identification. The DEVAN checklist by Vermeeren et al. (2002), utilised in this research (see Appendix B3), was developed specifically to detect usability problems in task-based products for adults. It provides a detailed list of behaviours that indicate usability problems. Zhao et al. (2012) employed this checklist in their study on the effect of different TA instructions on the outcome of CTA testing, and found that the checklist increases the reliability of the data collected. It should be noted that, following Jacobsen et al. (1998), Zhao et al. (2012) appended two further indicators to Vermeeren's checklist: "design suggestion" and "technical issues". These additional indicators were also used in the current research.

The application of a structured problem report: As Keenan et al. (1999) note, problem reports are often ambiguous, context-free, written in various styles, and of poor quality. This lack of clarity can lead to the inaccurate identification of problems (Cockton and Lavery, 1999). In contrast, structured problem reports encourage usability evaluators to carefully identify and analyse problems, which in turn increases the accuracy of problem extraction (Howarth et al., 2009). Capra (2006) conducted a detailed study on the elements that must be included in a usability problem report, and recommended the top five requirements:

- To be clear and precise while avoiding wordiness and jargon;
- To describe the causes of the problem;
- To describe observed user actions;
- To support findings with data;
- To describe the impact and severity of the problem;

These requirements are in accordance with the structured report form devised by Lavery et al. (1997). The current research adopts Lavery et al.'s (1997) format, which was specifically designed to standardise the process of usability problem extraction. The process includes the documentation of context, the framing of problems in terms of user difficulty and associated causes, and an examination of the impacts of usability problems on the performance of the participants (see Appendix B4 and B5).

Clear problem matching process: Law and Hvannberg (2008) described the process of matching problems (or consolidating problems) as involving the steps of problem extraction and problem filtering and merging, which can be done individually or collaboratively by evaluators. Hornbaek and Frøkjær (2008) warns that matching usability problem descriptions is not straightforward, but a difficult activity. In this regard, Lavery et al. (1997) and Hornbaek (2010) recommended the use of a structured report as a way to strengthen the process of problem matching. In this research, duplicated usability problems were merged to form a single problem if they rose from similar context and had similar descriptions.

The process of the usability problem identification in this research consists of two stages (Figure 3.6). In Stage One (*Individual probl*ems) each participant's testing video was

reviewed in order to detect usability problems. Data files were selected using a random number generator to reduce order effect. The usability problem indicators, were used at this stage to guide the extraction process. Each problem that was discovered was assigned a number (e.g., IUP1), and was recorded in a report in terms of the contexts in which they arose, their descriptions, their impact, their persistence (the number of times a problem is encountered by the same participant), the current task, and the time when it occurred (generated by screen capture recorder) (see Appendix B4).

In Stage Two (*Final problems*), starting with participant one, individual problems were merged across participants to form a final usability problem if they had similar problem descriptions and contexts. Structured reports were also used at this stage to record detailed information relating to each final problem (see Appendix B5). Each final problem was assigned a unique number (e.g. FUP1). All previous documents, namely individual problem reports, were attached to this final report.

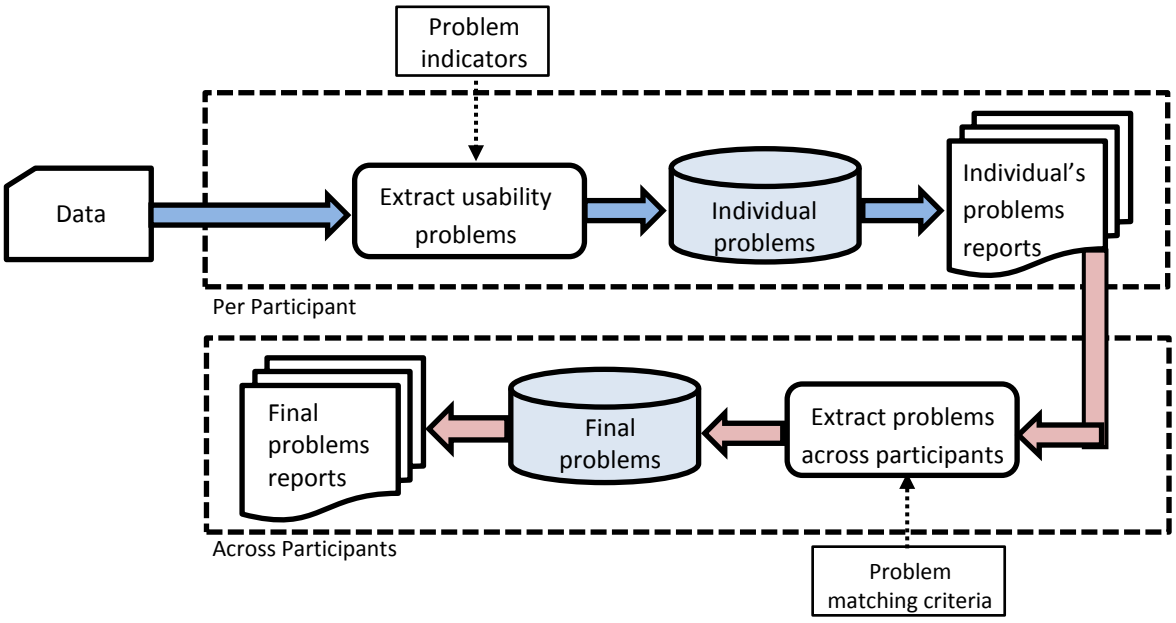


Figure 3.6: Schematic overview of the usability problems extraction process

Use of extra evaluators and coders: Hertzum and Jacobsen (2001) and Jacobsen et al. (1998) recommend the involvement of an additional evaluator to extract usability problems from the entire set of test data as a means of reducing the evaluator effect. However, such an approach demands considerable time and resources and is therefore very difficult for

researchers to implement (Barendregt et al., 2006). In the present research, full involvement of an extra evaluator for test sessions was impractical, so a trade-off approach similar to that of Barendregt et al. (2006) was employed by occasionally involving an additional evaluator to check the reliability of problem extraction. Furthermore, two usability experts were recruited in each study to divide all detected problems into specific problem types.

3.11 Statistical Analysis

In this research, two types of statistical analysis are used: descriptive and inferential[4]. Descriptive statistics (e.g. frequencies, central tendencies or dispersions) are the easiest to analyse: their primary purpose is to describe and summarise data so they can be easily understood and interpreted. They also intend to check data quality and aid in examining the assumptions of inferential tests.

Inferential statistics (e.g., t-test, ANOVA or Mann-Whitney) are used to identify relationship between variables, and to confirm whether conclusions regarding differences between levels of independent variables are valid and not merely due to random variation.[5] In HCI, inferential statistics are most often based on null-hypothesis significance testing (NHST). The NHST approach states a null hypothesis, which assumes no difference between conditions, and uses particular inferential tests as evidence for an alternative hypothesis, which assumes a significant difference (Hornbæk, 2011).

Determining whether an inferential test belongs to a parametric or non-parametric test group depends on the aim of the test, the design of the test, and the type of measurements of variables. Typically, non-parametric tests are used to assess categorical data, whereas parametric statistical tests are preferable for continuous data since they are more powerful than non-parametric alternatives. Parametric statistical techniques also hold some assumptions about the data such as the distribution of the data from dependent variable(s) is normal and that the homogeneity of variances is equal. If the data does not meet parametric test prerequisites, one can either use an alternative non-parametric tests, or

[4] Leading organizations are increasing their reliance on statistically significant data within their business decision making processes (Pyzdek, 2003)

[5] Random errors, also called 'chance errors', occur by chance and are not correlated with actual value.

manipulate the data which fail to satisfy the underlying assumption, or proceed with the analysis even when the data violates certain assumptions (Field, 2009)[6].

For the purposes of statistical analysis, all data were first transferred into Microsoft Excel for preparation and then into the IBM Statistical Package for Social Science (SPSS 22) for in-depth analysis. For ease of analysis, the format of the data allotted one participant to each row, a different variable to each column, and all variables to the same file. The names of variables were uniquely defined and were as descriptive as possible, and their types (e.g. categorical or continuous[7]) were clearly specified according to the types of values entered for those variables.

The statistical analysis process was undertaken at three levels, beginning with a separate analysis of each individual method within each single study. This was followed by a comparison between the TA methods in each study in order to reveal each method's relative performance. Finally, the researcher compared the performance of methods across the studies. For the sake of clarity, obtained values are presented in tables and figures. The results are then discussed, in the context of the currently available literature and the research questions posed.

3.12 Validity and Reliability

There are two crucial aspects of research methodology that any researcher planning and executing a study seek to maximise: validity and reliability. These are particularly significant in comparison studies of UEMs (Gray and Salzman, 1998). Validity can be defined as the degree to which "a study measures what it intends to measure", while reliability is a question of whether the same results would be obtained if the study were to be repeated (Krathwohl, 1997).

Gray and Salzman's (1998) commentary on five influential experimental studies comparing the usability tests, cognitive walkthrough, and heuristic evaluations, found that

[6] Laerd (https://statistics.laerd.com) and Usablestats (https://www.usablestats.com) provide useful statistical guides for novice researchers.

[7] A variable can be treated as categorical when its values can fall into specific categories (e.g. different educational levels) and as continuous when it can possess any value between two numbers (e.g. time on task).

these studies were severely lacking in their validity and therefore produced questionable results and conclusions. Based on these findings, Gray and Salzman provided recommendations for addressing the types of validity most relevant to usability evaluation research: internal validity, construct validity, statistical validity, external validity, and conclusion validity. These measures of validity, and the ways in which they were incorporated into the design of the current research, are discussed in more detail below:

3.12.1 Internal Validity

Internal validity refers to the level of confidence in the design of the experiment, the data collected, and the cause-and-effect statements that emerge from the study. It primarily seeks to verify whether the independent variable caused the observed change in the dependent variable or whether both variables simply correlated and a third unknown variable was responsible for the changes observed. While there is unfortunately no direct measure of internal validity, Gray and Salzman (1998) state that internal validity can be guaranteed through taking into consideration instrumentation, selection of participants and setting.

Instrumentation concerns evaluators' biases in identifying or rating the severity of usability problems. In the case of comparing UEMs, instrumentation is only valid if there is a systematic way of extracting and rating the severity of usability issues that does not inappropriately favour one condition over others (Gray and Salzman, 1998). In the present research, the same extraction and rating approach was employed in all three studies in order to reduce the impact of the evaluator effect and to maximise internal validity.

Selection concerns the characteristics of participants. There are two types of issues with regard to selection: general and specific selection threats (Gray and Salzman, 1998). A general selection threat occurs when participants' characteristics are not directly related to the manipulation under study, whereas a specific selection threat exists when participants assigned to different groups are unequal in terms of some characteristics (e.g. knowledge and expertise) that are directly linked to experimental conditions. To guard against both types of threats in this research, the researcher ensured that the recruited participants were potential users of the tested systems and that there was as much homogeneity as possible between groups.

Setting refers to the location and environment of an experiment. As Gray and Salzman (1998) note, variance in settings can threaten the internal validity of usability evaluation studies due to the difficulty of determining whether the effect observed in the study was obtained from the treatment, the setting or the combination thereof. Hence, all participants in each study in the current research were tested in the same physical location and environment to ensure accuracy of results.

3.12.2 Construct Validity

Construct validity considers whether researchers are in fact measuring what they claim to be measuring. To ensure construct validity, researchers should provide explicit information regarding the exact methods and procedures used so that readers will possess sufficient understanding to apply the same methods and procedures. It is also strongly recommended not to use the same participants for multiple UEMs in order to avoid the possible threat of interactions of different treatments, wherein participants' experience gained under method A may affect their behaviour under method B (Gray and Salzman, 1998). To take construct validity into serious consideration when undertaking this research, each method was clearly described and the variables and measures used were unambiguously and precisely defined. The problem of interactions was eliminated since the between-subject approach was used in each individual study, exposing each group of participants to only one TA condition, as explained earlier in section 3.5.3.

3.12.3 Statistical Validity

Statistical validity seeks to determine if there are significant differences between outcomes (dependent variables) in UEM groups, using one or more of a range of formal statistical techniques. The most common threats to this kind of validity include low statistical power and the insufficient use of established statistics. The statistical power of an experiment refers to the "probability of correctly rejecting the null hypothesis (i.e. no difference between groups) when it is false" (Gray and Salzman, 1998). Due to their small sample size, experiments with low power have a higher probability of incorrectly accepting the null hypothesis. The second threat to statistical validity lies in the fact that many UEM researchers tend to rely on simple descriptive statistics and "eyeball testing" rather than more sophisticated statistical tests such as inferential tests when deciding whether apparent

differences are significant (Gray and Salzman, 1998).[8] These two issues can be regarded as two sides of the same coin. Low statistical power may cause true differences to go unnoticed, which is known as a false negative or type II error; insufficient use of established statistics may mean that the differences that are noticed are not true, which is referred to as a false positive or type I error. Most problems with statistical validity can be avoided by using a large sample size (Gray and Salzman, 1998). To ensure that this research could obtain statistically valid results, the sample size in each TA group was large enough to accommodate the effect of low statistical power and allow for statistical validation analysis (inferential statistics).

3.12.4 External Validity

External validity refers to the extent to which findings in a study can be generalized to wider populations, settings, and conditions (Maxwell, 2005). Although generalisation can threaten external validity, this can be remedied by balancing grand claims against explicitly stated limitations (Gray and Salzman, 1998). In this thesis, external validity was achieved by explicitly stating the scope of the research and the possible limitations of the findings in the concluding chapter (Chapter 8).

3.12.5 Conclusion Validity

Any conclusions regarding a study must be drawn directly from the results of a study and supported by a chain of evidences (Gray and Salzman, 1998). For instance, ACM's CHI conference instructs authors that "the validity of your submission's contribution must be adequately supported by appropriate arguments, analyses, evaluations, or data as best fit the contribution type".[9] A study conclusion is considered invalid if the study claims are not investigated in the study or the data presented in the study contradicts these claims. In this study, all conclusions were drawn from the results of the study and were supported by descriptive and inferential evidences, and any speculated implications are clearly stated as being the opinion of the researcher. Table 3.5 below briefly summarises issues of validity and the solutions adopted by this research. These issues are discussed further in subsequent chapters of this thesis.

[8] Eyeball test refers to the practice of looking at the data and deciding by intuition that differences between tested samples are real.
[9] http://chi2013.acm.org/authors/guides/guide-to-a-successful-archive-submission/.

Table 3.5: Validity issues and resolutions

Validity Issue	Solutions
Internal Validity	- Avoided instrumentation problems by using a unified way to extract and rate usability problems - Avoided selection problems by ensuring that the recruited samples were as representative and homogenous as possible - Avoided setting problems by keeping the test location and environment consistent for all participants
Construct Validity	- Described clearly each method used and the exact procedure - Exposed each sample group to one TA condition only
Statistical Validity	- Provided a large enough statistical sample of participants
External Validity	- Ensured that findings could be easily generalised and replicated, by clearly describing the scope and limitations of finding, and what variables need to be controlled
Conclusion Validity	- Careful writing - Explicitly stated statistical data when quoting experience-based claims and stated any assumptions clearly

Regarding reliability, according to Hornbaek (2010) the most crucial factor in the reliability of the results of a usability study is the evaluator effect, which is most visible in the problem extraction process and which must be controlled in order to ensure the reliability of data. This is addressed in the context of this research by applying a number of measures to reduce the evaluator effect, as addressed in section 3.10.

3.13 Ethical Considerations

Ethical considerations are paramount in research, particularly when human participants are involved. In the context of research, ethics "define what is or is not legal to do, or what moral research procedure involves" (Newman, 2003, p.19). Factors that may give rise to ethical issues include the nature of the research project and participants; data collection procedures; the type of data collected; and the use and publication of data (Cohen, Manion and Morrison, 2007). The present research intends to follow the four standards of good practice: (1) doing positive good, (2) non-malfeasance, (3) informed consent, and (4) assurance of confidentiality and anonymity (Bošnjak, 2001). Ethical issues were not anticipated, as this research does not involve sensitive topics, participation from differently abled or vulnerable participants, and/or covert observation techniques.

Prior to data collection, the three empirical studies comprising this research were granted full ethical approval by the UEA ethics committee (Appendices C1, D1 and E1). In obtaining ethical approval, a pre-specified protocol was set out and agreed for each study, with all subsequent amendments to the protocols resubmitted to and approved by the

committee. The researcher also completed a data protection course at the University in order to better meet the university's data protection requirements.

Prior to each study, a full description of the purpose of the study and what it involved was given to all prospective participants through a recruiting script in order to enable them to make an informed decision regarding whether to participate. It was clearly explained to participants that their participation was voluntary, that they could withdraw at any time and without penalty, and that the observation sessions would be recorded. Participants had the opportunity to ask any questions they had about the study. If they decided to participate, their inclusion in the study was contingent upon providing a signed informed consent form allowing the researcher to use the data gathered from their participation as part of the study. The form also stated that participants could ask to view all work arising from the study, including this thesis.

Participants were guaranteed that data would remain confidential and would not be disclosed under any circumstances. Specifically, they were informed that data collected during and produced from the study would be stored in accordance with the UEA's data protection policy, compliant with the UK Data Protection Act 1998 (DPA)[10], which stipulates that for safeguarding purposes, personal information must be stored separately from other data and deleted when no longer needed. Accordingly, the consent forms to which participant signed their names were filed separately and destroyed within two months of the study session. Participants were further assured that the hard copies of the data would be stored in a locked cabinet and the soft copies on a password-protected computer in accordance with University policy.

During the study, participants were not exposed to any physical or emotional risk or harm beyond what could reasonably be expected to arise from the daily personal use of computers. Therefore, no additional safety measures were considered in advance. Given that this research focuses on participants' views regarding the system and TA methods under study rather than on individual names, and to protect participants' privacy, the researcher substituted numbers for participants' names in this thesis.

[10] The DPA may be found on the internet at http://www.legislation.gov.uk/ukpga/1998/29/contents

3.14 Summary

This chapter discussed and justified the choice of research methodology for the study and set out its various components. It provided the details of the research method, design, experimental procedures and data collection techniques used, and the strategies applied to analyse the collected data. The next chapter explores the use and effectiveness of classic TA methods in research.

4

CLASSIC THINK-ALOUD STUDY

4.1 Overview

This chapter presents the first empirical study. This study explores the impact of classic think-aloud methods developed by Ericsson and Simon (1993), namely the concurrent think-aloud, the retrospective think-aloud, and the hybrid on the outcome of usability testing. The chapter starts by defining the specific aims of the current study, identifies the tested online library, and outlines the test tasks and participants. Following this, it discusses the material and equipment used in the experiment, explains the experimental procedure, and presents the results of the pilot and main experiments. Finally, the chapter concludes by discussing and summarizing the results of the study.

4.2 Study Aims

This study aimed to investigate the utility and validity of the classic think-aloud (TA) methods, the concurrent think-aloud (CTA), the retrospective think-aloud (RTA), and the hybrid (HB), within the context of usability testing. Specifically, it examined the extent to which these methods can affect participants' task performance, their testing experience, and the usability problems discovered. Furthermore, the study explored the cost of employing the methods, and the relationship between the sample size and the number of problems detected in each condition.

4.3 Test Object

As mentioned in section 3.7, the researcher decided to use a university library website as a test object for the experiment in this study due to the growing popularity and widespread use of academic digital libraries, and the scant research that investigates the impact of TA methods on usability testing for such media. The selection of the specific university library website was based on the following criteria:

1. It had to be a dynamic website with multiple interactive features and functions.
2. It had to be manageable in size to allow for thorough evaluation of its usability level.
3. It had to possess a certain number of potential usability problems, thereby ensuring to some extent that participants would encounter difficulties whilst using the site. This would systematically be determined by conducting a preliminary heuristic evaluation of the potential site. More details of this are given in section 4.4.

4. To retain the validity of the results obtained, the interface of the selected site could not change after the heuristic evaluation or during the study period. This would need to be checked with the chosen website's administrator.
5. To ensure to the greatest extent possible that study participants could not rely on pre-existing knowledge of the website interface when performing test tasks, the site selected should be unfamiliar to study participants. If participants were frequent users of the chosen website, they could already expect to find certain types of problems and miss true usability problems (Sova and Nielsen, 2003). Moreover, they might not partake in the TA protocol to a sufficiently high degree and complete actions too quickly due to their expert status (Nielsen, 2010). This excluded well-known academic library websites such as the British Library website, as well as the University of East Anglia (UEA) library website, since the study took place at this institution.

After a careful evaluation of several websites, the University of East London (UEL) library website[11] was deemed a promising candidate for this study (see Figure 4.1). Once the website was selected, the researcher contacted the website administrator via email (see Appendix C2) to obtain consent to use the site in order to ensure the study's adherence to an ethical code of conduct, and to establish in advance that there was no intention to modify or alter the interface, either prior to, or during the study. Luckily, the administrator of the UEL library website gave the researcher written consent (see Appendix C3) to evaluate their website and informed him that the interface would not be modified prior to, or for the duration of the intended study period.

At the onset of the planning stage of this study in July 2013, the UEL library website, hereafter called UEL-L, had been serving as a portal to the library services and resources for six years. As shown in Figure 4.1, the UEL-L home page has a search engine positioned in the middle and a number of links for various options that are standard to most academic libraries' websites, such as conducting searches, booking a study room and booking a library PC. The website had a mixed base interface combining navigation and reading. All information on the site was only available in English.

[11] http://www.uel.ac.uk/lls/

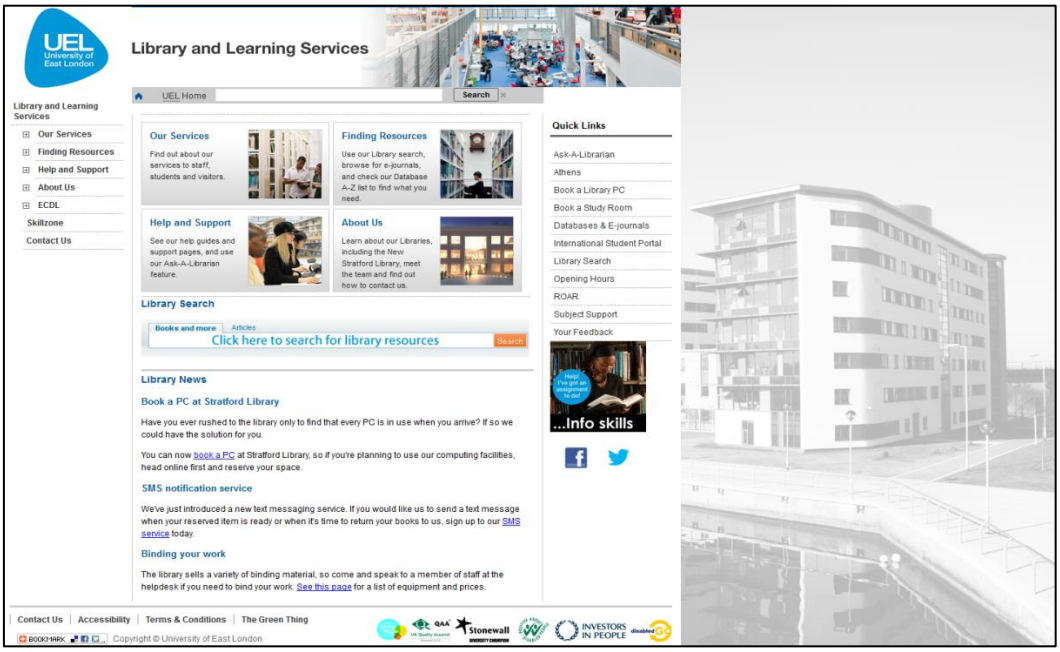


Figure 4.1: Screenshot of the test object's homepage

After defining the test object, a set of tasks was developed to assess the usability of the chosen website by means of the three TA methods.

4.4 Tasks

All thinking aloud usability tests involve the selection of a set of target tasks for the participants to perform during the testing session. It is impossible to test all the tasks that end users will do in a real situation. There are two important criteria for selecting tasks: 1) selecting those that are representative of the actual activities the end users undertake most whilst using the targeted application in a real-life context, and 2) selecting those that could be diagnostic in revealing usability problems (Dumas and Redish, 1999).

One way to ensure representativeness is to use tasks derived from an empirical investigation of users' activities rather than evaluators' fallible intuitions (Dumas and Redish, 1999). To this end, a context of use analysis of the chosen website was conducted. Context of use analysis is a generic method whereby the salient characteristics of the application under study can be determined by collecting and analyzing detailed information regarding users' characteristics, tasks and working environments. This process helps to mitigate the artificial nature of the evaluation process and improves the contextual validity of the results obtained. It also assists in identifying the limitations of an evaluation

and allows judgements to be made concerning the generalisability of that evaluation to other contexts (Maguire, 2001).

The application of this valuable analysis can take several forms, the most cost-effective of which is to interview stakeholders who have appropriate knowledge of the system under evaluation. Stakeholders may include, but are not limited to, project managers, trainers, developers and system support team members (Maguire, 2001). To obtain the information required for the present study, a structured interview with the website administrator was conducted. Wisker (2001) classifies the most common types of interview as follows: structured interviews, which involve a predetermined set of questions; semi-structured interviews, where the interviewer has worked out a set of questions in advance, but is free to modify their order based upon his/her perception of what seems most appropriate in context; and unstructured interviews, wherein the interviewer lets the conversation develop spontaneously and informally within a general area of interest. As Cohen, Manion and Morrison (2007, p. 355) state, "the structured interview is useful when researchers are aware of what they do not know and therefore are in a position to frame questions that will supply the knowledge required". Since the author was already aware of the information that needed to be obtained, a structured interview was deemed most suitable for this study.

Prior to the interview, an interview agenda was sent to the interviewee to adhere to an ethical code of conduct (see Appendix C4). This agenda outlined the purpose of the interview, the interviewee's rights and the confidentiality of the data, as well as the time and the estimated length of the interview. The time limit of the interview was set at 30 minutes in order to maintain active conversation.

Despite the structured nature of the interview, the interview guide included open questions which allowed the website administrator to provide additional responses or elaboration as necessary (illustrated in Table 4.1 below). This guide was made in advance by the author after surveying the relevant literature relating to context of use analysis (Bevan and Macleod, 1994; Maguire, 2001) and was reviewed by a native English proof-reader in order to detect weaknesses and to clarify ambiguity so that the interviewee could give his responses without experiencing any difficulty.

Table 4.1: Interview guide

- What is the purpose of your website?
- Have you done any recent user research on the site? If yes, explain
- Who are the primary users of the site?
- Are there any secondary users of the site? If yes, who are they?
- Would you describe the users (primary and secondary) of the site? (Age, gender, education level, web experience, nationality, mother tongue, physical and sensory ability, etc.)
- What tasks do the users perform the most frequently on the site?
- Which tasks are most important to the users (primary and secondary)?
- How do the users access the site? (Via desktop/laptop browser, mobile browser or both)
- Are there any other contextual factors that might affect the user experience? If yes, explain.
- Are there any problematic areas or design issues in the website? If yes, please explain.
- Would you like to add anything else? If yes, what would you like to say?

At the interviewee's request, the interview was conducted over the phone. The interviewer (author) opted to use a mobile phone instead of a landline phone to benefit from the speaker tool, which enabled clear recording. Before starting the telephone interview, the interviewer reassured the interviewee that he could withdraw his participation at any time without repercussions. The interviewee was also informed that the interview would be recorded using an Olympus device[12] (Version, 2013) and gave his oral consent for the recording to take place. The researcher's experience of using the Olympus device in his Master's research project confirmed its robust practicality. During the interview, the researcher began recording and taking notes. The web administrator voluntarily offered some information regarding the website audience and the usage information on the site. The interview lasted approximately 18 minutes. After the interview, a follow-up phone call was made to clarify certain points.

Table 4.2 in the next section summarises the information gained from the process set out above. The administrator stated that, based on usage statistics, users of the library's website mainly accessed the site to search the library catalogue using a variety of search options; find out about "Athens" to access resources off-campus; check contact details of their subject librarian; ask reference questions via the "Ask a librarian" service; find out about library services and updates; and look up hours of operation for the library.

As noted in the previous section, the library site was evaluated thoroughly during the planning stage of this study by the researcher using the heuristic evaluation method (heuristic evaluation is described in chapter two) in order to identify potential usability

[12] http://www.olympusamerica.com/cpg_section/cpg_support_manuals.asp?id=1658

problems which, in turn, could provide a focus for the task design. In addition, a usability expert evaluated the selected website in order to further confirm the results. The heuristics evaluation, which was based on the widely used heuristics principles developed by Nielsen (2000) (see Appendix A), found that the library site possessed a number of predicted usability problems varying in nature and severity and was thus a suitable object for the study. Most of the usability problems predictions detected were related to four heuristics: visibility of system status, user control and freedom, error prevention and correction, and aesthetic and minimalist design. Examples of these problems included ineffective internal search functions, text that was highlighted on roll-over but was not clickable, use of too many hyperlinks, and ambiguous links. The author utilised the results of the heuristic evaluation alongside the information acquired from the website administrator regarding users' activity patterns on the site to guide the design of various tasks (see Figure 4.2).

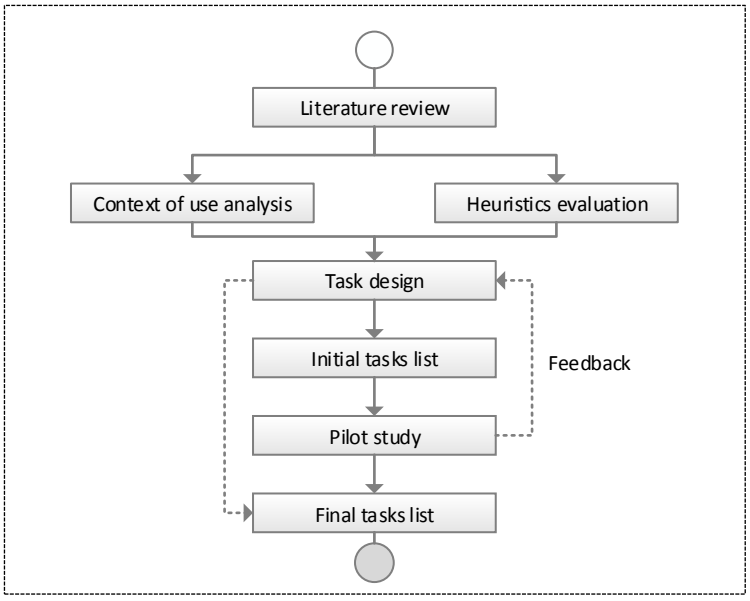


Figure 4.2: Tasks development process

Seven tasks were designed that together covered the targeted website's main features and predicted problematic areas. These tasks were intended to be neither too difficult nor too simple, as both extremes might prevent participants from verbalising and would negatively affect the time required to carry out the tasks (Ericsson and Simon, 1993).

The constructed tasks each focused on a different area of the site to the avoid learning effect, and were designed to be completely independent of each other so that failure in one task does not impact the overall process. Task one evaluated the ease of navigating the site

to find the name of a subject support. Task two assessed the booking function for study rooms on the site. Tasks three and four evaluated the site catalogue's "simple search", while tasks five and six evaluated the catalogue's "advanced search" and "sort results" functions. Finally, task seven examined how participants worked with viewing search history on the site (see Appendix C5).

Each task was presented in a scenario format. Scenario formats are the most widely used task formats in usability testing, as they facilitate the emulation of real-world contexts of use and thus enhance the ecological validity of tasks (Dumas and Redish, 1999). The written description of each task scenario clearly set out what participants were expected to try to achieve and their motivation for doing so, and was as short as possible to keep the testing session moving quickly. The tasks required the participants to begin at the homepage then navigate through the website attempting to *"find"* a particular piece of information that was known to exist on the UEL-L site. There was only one correct answer per task, which allowed both the researcher and the participants to determine whether the task was completed successfully. It was important that the tasks did not request common information that the participants might already know without having to use the interface. Additionally, the search topics for tasks three to six had to avoid causing offence to any participants, be gender-neutral and revolve around topics that were generally familiar.

As shown in Figure 4.2, all tasks were subjected to three pilot tests prior to final testing to ensure that they were free from bias and ambiguity and were sufficiently different from one another, as well as to estimate the average time required for completing them. In pilot testing, the seven tasks took an average of 20 minutes to complete, which was considered to be an appropriate length of time. Longer completion times could have led to participants becoming restless and impatient, while shorter completion times could have prevented participants from sufficiently familiarising themselves with the testing process. Following pilot testing, the finalised task list was created (see Appendix C5). An example task is shown below:

> 'Task #4: You want to find the journal paper that has the title "Building for the Future" written by Doyle Henry in 1963 to read before a coming seminar in education subject. Can you find it?'

Once the experimental tasks were constructed, the recruitment of participants commenced. The next section discusses this process in more depth.

4.5 Participants

The recruitment of study participants requires careful thought and effective planning as the quality of this process will have an immediate impact on the quality of the data collected. Sova and Nielsen (2003) assert that there are four steps which must be followed prior to the recruitment process in order to avoid compromising the validity of the results:

(1) To include the right sample size;
(2) To learn about the user profile;
(3) To develop clear and precise recruiting criteria;
(4) To determine the appropriate recruiting method.

Sample size

As seen in the literature review in the second chapter, the question of what constitutes an optimal number of participants for a usability test is one of the most heated debates in the field. Some researchers state that five to nine participants are sufficient for an effective usability test (Nielsen, 2000; Nielsen and Landauer; 1993b). However, these numbers are arguably not applicable to the current study, as it aims to investigate the use of different TA usability testing methods rather than to detect usability issues using only a single method.

Although there is likewise little agreement regarding the optimal sample size for comparative usability studies, for this study it was decided that 20 participants would be recruited to each TA testing condition. This figure was based on the grounds that this study is not a typical stand-alone usability test where five to nine subjects are (controversially) adequate, but an experimental study of the relationships between independent and dependent variables which needs more participants to ensure statistical validity (Gray and Salzman, 1998). A sample size of 20 for each TA method creates sufficient statistical power to provide a stable estimate (Macefield, 2009), and reduce the "Wildcard effect", wherein a participant might have more or less than average experience with the type of

system being tested (i.e. library websites) (Gray and Salzman, 1998) [13]. Moreover, using 20 participants would ensure that the results of the study would be well suited to analysis with established statistical methods (Sauro and James, 2012), and is also very likely to produce statistically significant findings (Macefield, 2009). Furthermore, according to the 20/20 rule, there is typically a margin of error of approximately 20% in quantitative usability measures for 90% of the time with a sample size of 20, as opposed to a margin of error of 58% with a sample size of 5 for example (Sauro, 2010). Indeed, a number of between-subject TA studies were found to employ 20 participants for each TA condition (e.g. Van den Haak et al., 2004; Zhao et al., 2012; Olmsted-Hawala et al., 2010).

Following Sova and Nilesen's (2003) recommendations to devote some of the resources for any given usability study to pilot testing, the intent in this study was also to recruit three independent individuals to conduct the pilot study, and another three back-up participants to cover for inevitable cancellations or no-shows in order to ensure obtaining a full complement of participants for the main study. This made the desired sample size for all components of the study 66 participants.

User profile

As with tasks, the most important consideration for usability participants is that they are representative of the targeted user groups of the product being evaluated in order to provide the valid feedback needed to make meaningful improvements to a design. Non-representative participants are very likely to generate false problems and miss problems with the system that actual users will encounter. To obtain representative participants, the target audience of the system under evaluation must be understood so that an accurate user profile can be created (Sova and Nielsen, 2003). The context of analysis of the tested website permitted the researcher to identify the intended audience of the site. The site administrator indicated that the library site mainly caters, as expected, for students who are the dominant users of the site (85% of the site's users are students) and academic staff at UEL, although it can also be accessed by other staff and guests (i.e. people outside the university), who together represent its secondary users. The site allows its visitors to browse and access basic functionalities, except for some restricted areas such as 'loaning'

[13] Sauro and James (2012), under the headline '*Do You Need to Test at Least 30 Users?*', argued that the 'n=30 rule of thumb' has a very weak rationale in the context of user research.

or "reserving". Its interface is mainly accessible on desktops and laptops and serves a diverse student body with a wide range of academic levels, areas of study, and ages (18 years old and upward). However, it is not flexible enough to accommodate the needs of people who are differently abled, due to the small proportion of students with such needs at UEL. The majority of the primary users of the site were assumed to have "good web experience". The site administrator stated that there was a lack of information regarding the characteristics and demographics of the site's secondary users.

Table 4.2: Results of the context of use analysis

Primary users	**Main task goals**	**Personal attributes and skills**
Students Academic Staff	To search the catalogue To find out about 'Athens' To ask reference questions To check contact details of their subject librarian To find out about library services and news To look up library opening times	Male and female 18 years old upwards Different backgrounds (British are the majority) Undergraduates and postgraduates Different areas of study Regular Internet users Significant minority with disability
Secondary users	**Main task goals**	**Personal attributes and skills**
Other staff Visitors	To find out about services and news To look up library opening times To search the catalogue	Not specified

Recruiting criteria

After the user profile was clarified, a number of recruiting criteria were developed to obtain the most appropriate participants for the current study (see Table 4.3). These criteria which were derived from the context of use analysis and from TA and usability testing literature, were as follows:

1. Given the sample size of the TA groups in the current study, it was not possible to provide valid representation of different user subgroups. After discussing these difficulties with the supervisor at UEA and with other experts in the field, the researcher decided to select the study sample from among university students, as the site administrator deemed them the dominant and most important user group of the tested website. Targeting university students was also expected to facilitate the process of finding a truly representative sample of participants, which in turn would facilitate the selection of participants.

2. The participants sample must include male and female members, as the targeted site was intended for both genders.
3. The age range of the recruited participants should be 18 to 65 years old, the age was limited to 65 years old to limit the influence of ageing on TA usability testing (Sonderegger et al., 2016; Olmsted-Hawala and Jennifer, 2012)
4. Participants had to have competence in English due to the potential impact of language proficiency on TA behaviour (Sun and Shi, 2007).
5. Participants had to possess "good" Internet skills. Participants who had not previously and frequently used the Internet would spend most of their time attempting to master the basic technology required to access the Internet, and would be likely not to reveal anything significant concerning the usability of the system (Sova and Nielsen, 2003).
6. Participants could not have prior familiarity with the chosen website.
7. Due to the nature of the study, people with hearing, speech, sight, social/communication or dexterity disabilities were sadly outside the scope of the study (see Table 4.3).

Table 4.3: Recruiting criteria

Participants type	Students
Gender	Male and female
Age	18-65
English skills	Fluent
Internet use (excluding email)	Used the Internet at least once a week
Test website experience	Never used the selected website
Mental and sensory ability	No limitation in dexterity, speech, hearing, or sight

This type of sampling is known as purposive sampling technique, in which researchers identify and select individuals or groups based on predefined criteria. According to Denscombe (2007, p. 15), purposive sampling is most appropriate when "the researcher already knows something about the specific people or events and deliberately selects particular ones because they are seen as instances that are likely to produce the most valuable data".

Once the recruiting criteria were established, a screening questionnaire was created based on these criteria to ensure that all study participants were appropriately qualified. In order to maximise the effectiveness of the screener, the researcher consulted survey handbooks and reviewed relevant scientific papers (e.g. Brace, 2008).

The screener was divided into three sections as follows: Section one sought to gather information about the characteristics of the respondents and their background (e.g. mother tongue, age, gender, and nationality). Section two contained questions about their Internet and library websites background, and addressed their Internet browser to determine which one to use in the experiment. Section three covered their prior experience with thinking-aloud usability testing studies, and included a question about the candidates' willingness to have their voices and on-screen usage actions recorded during the experiment.

All the screening questions (closed and open questions) were written in such a way as to obscure which answers met the research criteria. For example, for the question related to candidates' experience with the selected website, the name of the test object was not given; instead, candidates were asked an indirect question: "Have you used any digital libraries before? If yes, please state which library website(s) you have used, starting with the most recent ones". This was also to prevent participants from preparing for the test in advance, which could have occurred if they knew which site was to be tested.

Prior to distribution, the screener was discussed with the researcher's supervisor, revised by a scholar in human-computer interaction and an English-language professional, and tested by three students who were chosen from the UEA student body. Students were approached individually and were given adequate time to complete the form. The piloting process sought answers to the following questions:

- Were all questions clear?
- Would students object to answering any of the questions?
- Did all questions yield usable data?
- How long, on average, were students likely to need to complete the screener?

The outcome of the above process revealed that the data collected was certainly usable and the students did not refuse to answer any questions. Based on the feedback received, the screening questions were further revised and compiled into the final model of the questionnaire, which is shown in Appendix C6.

Recruiting method

Generally speaking, there are two methods of recruiting usability study participants: agency recruiting and self-recruiting methods. In the former, an external agency is hired to find participants and make all necessary arrangements based on the study criteria, leaving the researcher free to focus on the study. Such a service can be especially helpful in recruiting a large quantity of participants in an efficient and convenient way, but the success of this method is contingent upon the quality of services provided by the agency (Sova and Nilesen, 2003).

To enquire about recruiting a representative sample of UEA students, the researcher emailed three recruiting agencies. A sample response received from one of the agencies is below:

> *"I'd be happy to give you a rough idea of the pricing based on the information you've provided so far. The range of pricing would likely be between £3500 and £4000. The total cost could be determined once we know the timeline of the study*."

Due to the large costs that this method would incur, the researcher opted for the self-recruiting method instead. This is probably the most widely used recruiting approach, primarily because self-recruiting allows for the recruitment of a wide range of participants at a low cost. In a survey by Sova and Nilesen (2003), 79% of usability practitioners indicated that usability participants are recruited by the practitioners themselves, rather than by an agency or other outside party. While self-recruiting enables the researcher to retain control over the recruitment process and the quality of participants recruited, this method requires strong project management skills and considerable time and effort (Rubin, 1994). Prior to undertaking the current study, the researcher had already conducted a number of small-scale usability studies in which the self-recruiting method was adopted (e.g., Alhadreti et al., 2011).

Participants were recruited from five sources:

- An email was circulated through official channels to students at UEA, Norwich, UK (see Appendix C7).
- A poster was displayed on departmental noticeboards at UEA at the UEA Sportspark (see Appendix C8).

- Social media networks such as UEA faculty Facebook pages were used.
- Several UEA instructors were contacted via email and asked to present the study to their classes and encourage any interested students to contact the researcher.
- The researcher's network of friends was employed.

When advertising for the study, the researcher endeavoured to establish mutual trust and rapport with prospective participants in order to overcome their potential misgivings. The screening script, therefore, outlined clearly all the information that prospective participants might need to know in order to decide whether to participate, including: the purpose of the study, the expected length of the experiment, the benefits of taking part, the level of risk involved if any, the study locale, data protection and anonymity, and the contact details of the researcher in case participants required further clarifications regarding the study.

The researcher avoided using the word "test" in the description of the study, as people generally feel anxious about tests (Sova and Nilesen, 2003); instead of referring to "usability testing", less intimidating phrases such as "website review" or "usability study" were used. It was also emphasised that the aim of the evaluation session was not to assess the subjects' skills or knowledge, but rather to evaluate the usability of a website interface, as recommended by Tullis and Albert (2008). To motivate more people to participate, a monetary incentive of £5 was promised as a token of appreciation for those who were chosen for the study.

Due to the value of online-based surveys in facilitating data gathering and analysis, a web-based tool called SurveyMonkey[14] (Version, 2013) was chosen to distribute the designed screener, making the instrument more environmentally friendly. SurveyMonkey provides real-time access to data to enable immediate and detailed analysis in the form of graphs, spreadsheet, and charts. A link to the screener was provided in the recruitment email and poster. A few copies of a paper-based version of the screener were prepared in case any participant prefer the traditional form; none of them were used.

A total of 102 screening questionnaires were completed by potential participants. These participants' answers were then screened to ensure that they fit the required profile, as set

[14] https://www.surveymonkey.co.uk/

out in Table 4.3. Of the 102 candidates who responded, sixty students meeting the selection criteria were contacted and invited via email (see Appendix C9) to participate in the study (see Table 4.4). 42 candidates were disqualified, as their demographic details did not meet the screening criteria or the required number of participants for the main study had been reached. As planned, three students on the excluded list were located for the pilot study, and another three students were invited as back-ups to offset no-shows.

Table 4.4: Distribution of potential participants

Potential	102
Excluded	42
Regular	60
Pilot	3
Backup	3
Total invited	66

The sixty volunteers recruited for the main study were allocated to the three TA testing conditions, with 20 per condition. To mitigate the impact of individual differences and to be able to draw valid comparisons between the TA groups, participants were matched on the basis of demographic variables as closely as possible. Participants with similar profiles were evenly assigned to the three testing groups in a matched randomised way, using a random number generator. Section 4.9.1 provides more details regarding the participants in the main study.

Once participants were assigned, they were asked to choose a convenient time for them to take part in the study from a set of predetermined time slots using a web-based scheduling tool called Doodle (Doodle, 2014). The pilot study was scheduled to be deployed over a two-day period from 15th October to 17th October 2013, and the main study over six weeks from 20th October to 4th December 2013, with two to three participants per day. Each participant was scheduled for a maximum 60-minute session. This timeframe was set based the researcher's experiences of conducting usability testing, and the researcher also did not think students would be able to commit for longer.

The evaluation sessions were arranged during weekdays, as people tend to be more reliable on weekdays and less reliable on weekends (Sova and Nilesen, 2003). Some sessions were scheduled during evenings for participants who could not attend sessions during regular

time hours. Moreover, a confirmation email was sent a few days before the scheduled sessions to proactively reduce the rate of no-shows (see Appendix C10). Figure 4.3 provides an overview of this process.

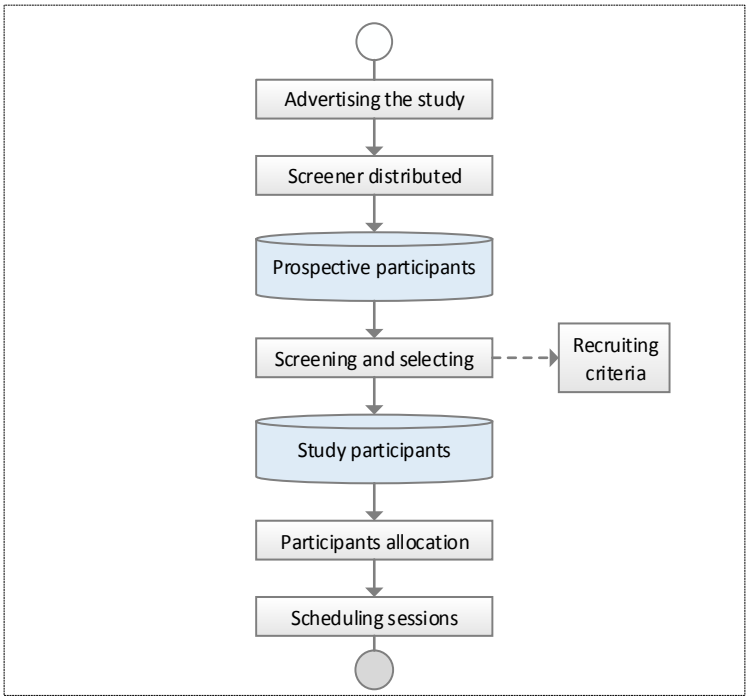


Figure 4.3: Recruitment process

It should be noted that although the researcher refers to participants in this thesis as "participants", "students", "subjects", "individuals", "users" and "volunteers" interchangeably, the term "participants" is used most frequently in order to highlight the collaborative nature of this research between the researcher and participants (Simons, 2010).

4.6 Setting and Equipment

All evaluation sessions were conducted in the same laboratory in the School of Computing sciences at UEA. An easy-to-follow map with clearly written directions to the test location was provided to participants with the invitation email (see Appendix C9). The test facility consisted of a waiting area and a testing lab with a comfortable table and chair, a standard personal computer for the participant, books on shelves, pin boards, posters, plants, and other items belonging to a typical office.

The environment and equipment were controlled to ensure good experimental practice and to reduce the chance of bias occurring due to participants having different equipment or surroundings (Lindgaard and Chattratichart, 2007). To ensure that the environment was comfortable for participants, noise levels were kept to a minimum with the ambient temperature within a normal range, and with appropriate lighting. Only the participant and the evaluator (author) were present during the experiment, which guaranteed participants' privacy. The evaluator was seated behind and to the right of participants to lessen the feeling of being observed and to minimise distraction. It was believed that the physical presence of the evaluator will make participants feel less self-conscious about thinking aloud. The setup of the lab, which is similar to typical practice (Rubin and Chisnell, 2008), is shown in Figure 4.4.

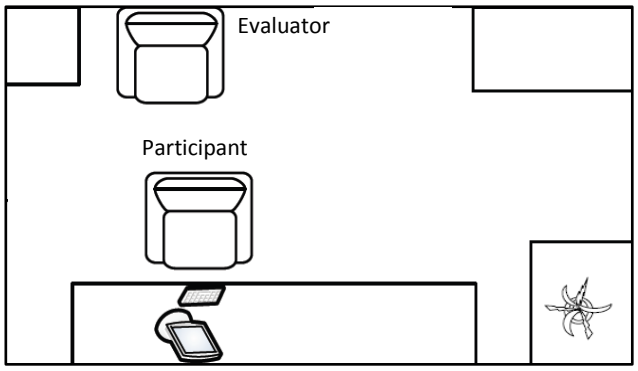


Figure 4.4: Setup of testing lab

In order to control for variation in computer performance, the same lab computer was used in all tests. The machine was equipped with Windows 7 Workstation (64 bit operating system), a GHz 2.80 Intel processor, a high-speed connection to Google Chrome, an external dual headset with a microphone, and a computer mouse (see Figure 4.5). The UEL-L website was set up as the default homepage in the browser. Google Chrome was chosen as the browser due to its widespread use in the UK in general (Statcounter, 2013)[15]and amongst the study participants in particular, 83% of whom were active Google Chrome users.

Camtasia software, running on the test machine, recorded participants' on-screen (keyboard and cursor) actions and verbal reports on video; participants' facial expressions

[15] http://gs.statcounter.com/

were not captured. It might be argued that including the face of participants could have provided more information about participants' emotions and opinions, but it also might concern participants with this issue, obscure a portion of the computer screen, and require evaluators to divide attention between three channels of information when analysing the data: think aloud (auditory), screen capture (visual), and participant's face (visual). A study by Long et al. (2005) compared two different versions of a digital usability session movie. One version had screen capture and think aloud, the other also had a video of the user's face in the bottom-right corner. Long et al. found no significant difference in the number of problems identified by each group, indicating that omitting the participants' faces does not impair problem detection.

A check list was developed to remind the evaluator to confirm before each session started that the test computer and all recording devices and equipment were fully functional, and to apply anti-bacterial wipes on the keyboard and mouse of the test laptop to help to protect participants from any possible infection (see Appendix C11).

Figure 4.5: Equipment used (picture taken with participant's permission)

4.7 Experimental Procedure

Prior to undertaking the study, permission for the experiments was sought and granted from the UEA Ethics committee (see Appendix C1). When participants arrived at the laboratory, they were cordially greeted by the evaluator and made to feel at ease. Participants were then asked to review and sign an informed consent form (see Appendix C12), which provided an overview of the study along with details of data storage and confidentiality. The form was written in plain, understandable language to avoid discouraging participants (Sova and Nilesen, 2003). The evaluator presented the study as

a typical usability evaluation: participants were informed that the purpose of the study was to evaluate the usability of a university library website, and not themselves. They were not told at this stage about the study's focus on TA methods (i.e. putative usability study), although this information was divulged to them at the end of the study. This was because if participants knew which TA condition (treatment) they were receiving and believed that it would affect the outcome, then the evaluator may have been measuring the effect of the belief rather than the effect of the treatment (confounding of belief in effectiveness of treatment with the treatment itself), which would affect the construct validity of the study. Participants were requested not to discuss the experiment with any other potential candidates and had the option to choose not to participate in the study after reading the consent form and/or to leave the study at any time without repercussion. When the participant signed the consent form and stated that s/he was happy to proceed, the evaluator moved on to the testing instructions. The respective protocol for each TA testing condition were set out in procedure instruction sheets issued by the evaluator (see Appendices C13, C14 and C15).

Two scenarios were initially proposed for applying the HB condition:

1. To ask participants to perform each task while thinking aloud and to give a retrospective report immediately after the completion of each task.
2. To ask participants to perform each task while thinking aloud and to give a retrospective report immediately after the completion of all tasks.

A problem with the first option was that inserting a retrospective account between individual tasks could have made participants more aware of being tested and thus produced biased results. The second option was deemed more suitable for this study, as collecting the retrospective protocols after the completion of all tasks would avoid the possibility of bias (see Appendix C15). Indeed, a single retrospective session appears to be the most common method in RTA testing (Leanne et al., 2016). The researcher noted Ericsson and Simon's (1993) caution that if a retrospective session follows multiple tasks, then the tasks need to be quite different to avoid participants generalising across episodes, as discussed in Section 4.4.

HB condition: In the concurrent phase of the HB condition, participants were first asked if they were right- or left-handed (for mouse configuration), and were given a maximum of two minutes to familiarise themselves with the test laptop and to regain their normal speed of interaction with computer systems. On completion of this step, the evaluator introduced the concept of thinking aloud using Ericsson and Simon's instructions (1993). Participants were instructed to think aloud while performing the tasks and to not turn to the evaluator for assistance; they were also informed that if they fell silent for a while, the evaluator would remind them to keep thinking aloud (see top row of Table 4.6). These instructions were followed by a brief TA practice session, as recommended by Ericsson and Simon (1993), in which participants were invited to practice thinking aloud using a simple, neutral task of looking up the word "*carol"* in an online dictionary (unrelated to the use of selected website) (see Figure 4.6).

After the practice session, the evaluator presented the task instructions sheet (see Appendix C16) to the participants, who were asked to read the instructions first to make sure they understood these fully before proceeding to task solving. Participants were instructed to complete the tasks in the sequence presented and to explain each task using their own words before starting to ensure that they understood the task requirements. To measure task completion times and status more accurately, they were asked to verbally alert the evaluator when they were ready to begin a task, and if they had found the necessary information or decided to give up the task. The evaluator did not indicate whether or not a task was successfully completed, and did not inform participants that they were being timed to avoid making them panic or feel they were being examined. The UEL-L website contains a major search feature, as seen in Figure 4.1, however, participants were encouraged to use this only if they felt they had no other choice, in order to encourage them to explore the website in more depth.

After introducing the test website and setting up the screen capture software (Camtasia), participants began to perform each task in turn. The tasks were displayed to participants on separate notecards in a counterbalanced order to prevent the order in which the tasks were presented from affecting the results (Sauro, 2010). This was achieved by counterbalancing the tasks sequence through the application of a Latin square: creating a grid of the number of tasks and the number of participants, then alternating starting tasks

by moving each successive starting task to the end of each successive row (Sauro, 2010) (see Table 4.5, Appendix C17).

Table 4.5: Sample order of task presentation

Participants	**Order of task**						
P1	1	2	3	4	5	6	7
P2	2	3	4	5	6	7	1
P3	3	4	5	6	7	1	2
P4	4	5	6	7	1	2	3
P5	5	6	7	1	2	3	4
Pn	..	..	..	..	..	..	..

During participants' tasks performance, the evaluator strictly followed Ericsson and Simon's (1993) guidance, and only issued a neutral TA reminder ('please keep talking') if the participants fell silent for 15 seconds; there were no other interactions. The evaluator tried to keep body language to a minimum at this stage.

While participants were working on each task, the evaluator recorded on an observation sheet the task completion status and time taken to complete the task (measured by a digital timer). Each participant's ID number and the date and start and end time of each session were also recorded (see Appendix C18). Participant names were replaced with participant IDs so that an individual's data cannot be tied back to individuals.

After all tasks were completed, the evaluator ended the recording and directed the participants to fill in the first online post-test questionnaire, the System Usability Scale (SUS) (see Appendix B2), to assess their satisfaction with the usability level of the tested website. Having done that, participants were then asked to complete the first two parts of the second post-experiment questionnaire (Experience with the TA Test), containing questions on their estimation of their method of working on the tasks compared to their normal working (part one), and their experience of thinking aloud (part two) in order to measure their testing experience as discussed earlier in section 3.9.2 (see Appendix B1). For each participant, the order of statements was randomised to limit the order effect and all items are evaluated on a 5-point Likert scale, ranging from 1 = strongly disagree to 5 = strongly agree. This phase was considered complete as soon as participants were finished.

Table 4.6: Concurrent and retrospective reporting instructions

Think-aloud phase	Instructions
Concurrent phase	In this study, I am interested in what you say to yourself as you perform some tasks that I give you. In order to do this I will ask you to **think aloud** as you work on the tasks. What I mean by think-aloud is that I want you to say out loud everything that you would normally say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?
Retrospective phase	Now that you have finished the tasks, I would like you to watch your recorded tasks performance on muted video and give **retrospective reporting** on them. In other words, I would like you to recall the thoughts you had when completing each task, and tell me any thoughts you had. Do you understand what I want you to do?

Once the concurrent phase was complete, the evaluator introduced the retrospective phase using Ericsson and Simon's (1993) instructions (see bottom row of Table 4.6. Participants were asked to watch their recorded performance on muted video and give retrospective reporting. The video showed RTA subjects pages visited while doing the tasks, the cursor movements, and the keyboard actions made. The use of video recordings as a stimulus for RTA is documented in the existing literature (e.g., Van den Haak et al., 2004; Peute et al., 2010). During this phase, the evaluator did not intervene, apart from reminding participants to think aloud if they stopped verbalising for 15 seconds. This separation was fundamental in reducing the possibility of the evaluator unwittingly biasing the data collected or participants' responses to the evaluator's questions or prompts (Ericsson and Simon, 1993). Camtasia recorded the retrospective verbalisations of participants reviewing their task behaviour. Upon completion, the questions posed in the second part of the TA testing experience questionnaire regarding the experience of having to TA were repeated after the retrospective phase in order to investigate whether participants would have different experiences of thinking aloud after the retrospective stage. Afterwards, the participants filled in the third part of the participants' testing experience questionnaire (evaluator presence), including questions on their opinions regarding the presence of the evaluator.

CTA condition: The instructions and procedure for the CTA condition were exactly the same as for the concurrent phase in the HB condition (see Appendix C13). However, participants in the CTA condition filled in all parts of the post-experiment questionnaires at the very end of the experiment.

RTA condition: In the RTA condition, the evaluator first instructed participants to familiarise themselves with the laptop and perform the preliminary task. They were subsequently asked to review the task instruction sheet and then to solve the seven tasks in silence without the assistance of the evaluator. During testing, the evaluator observed and took notes, but did not interact with participants. As with the HB condition, the retrospective protocol in the RTA condition was collected after the completion of all tasks rather than after each individual task in order to reduce the possible impact of individual retrospective accounts on subsequent tasks. At the end of the final task, the participants were asked to fill in the SUS questionnaire, and the first part of the Experience with the TA Test questionnaire. They were then instructed to voice their thoughts retrospectively while watching muted videos of their actions. The instruction for this stage was exactly the same as for the retrospective phase in the HB condition (see Appendix C14). Subjects were then able to practice thinking aloud. After completing the retrospective reporting, participants were directed to fill in the remaining parts of the Experience with the TA Test questionnaire.

After the session concluded and the evaluator checked that all required documents had been filled out, the evaluator thanked each participant for taking part and gave them their monetary honorarium in an envelope labelled with their name - providing the incentive at the end of the session ending the session on a positive note and minimised the sense of obligation to speak positively (Sova and Nilesen, 2003). Participants then signed a receipt indicting that they had received the compensation and left. Following that, all documents and notes related to each participant's testing process were collated, and video footage of the participant's screen actions with his/her voice was compressed and copied to a folder identified by the number assigned to the participant. Finally, the testing environment was restored to its original condition in preparation for the next experiment, and all the search history on the site were deleted, so the next participant got to start from scratch. Figure 4.6 below depicts the experimental procedure for the three conditions.

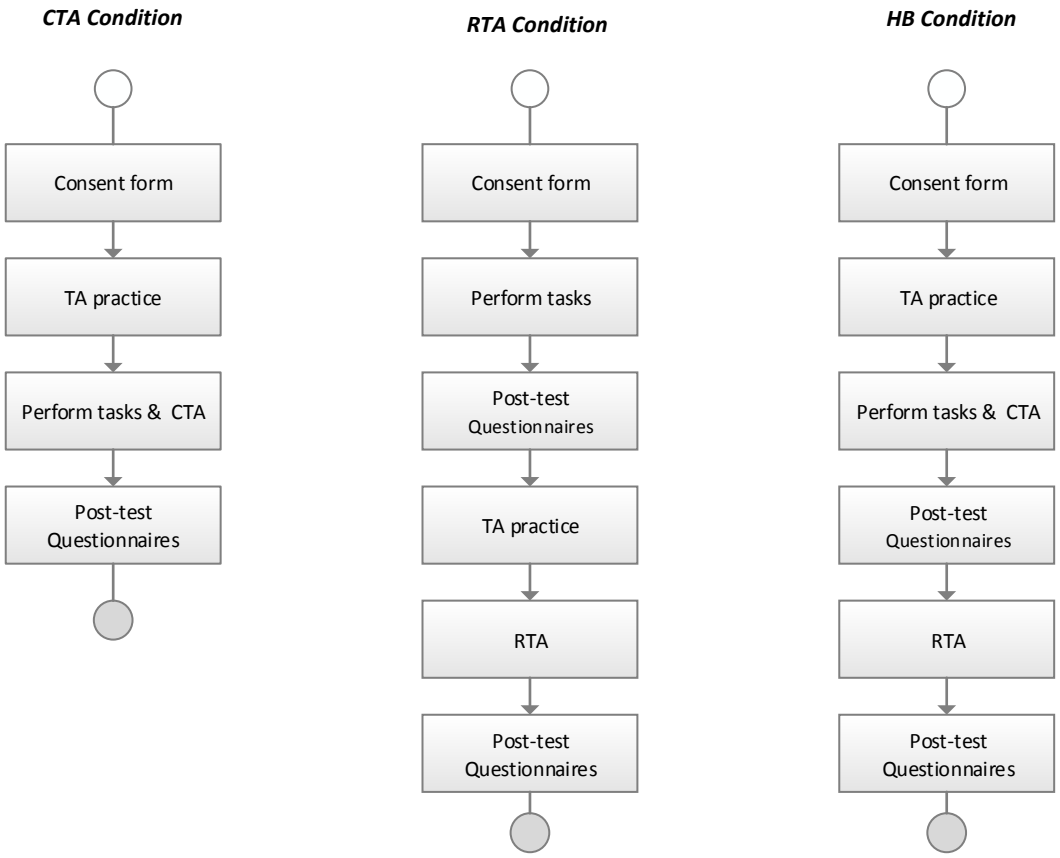


Figure 4.6: Experimental procedure

4.8 Piloting and Correction

Prior to the actual study, a pilot study took place to ensure that the experiments were effectively planned. Conducting the pilot study allowed the evaluator to review and fine-tune the experimental system, focusing specifically on the methods used for data collection, the accuracy of documentation and the effectiveness of the tasks list.

As detailed earlier, the pilot study was individually conducted with three UEA students. One participant was assigned to each TA testing condition. It took place in the same usability lab and largely under similar conditions as the actual experiment, i.e. machine used, Internet connection, type of browser, and surroundings. Two subjects in this experiment were British male students and one was a Saudi male student. They had no further involvement with the main study, and their data were not included in the raw data set of the actual study.

During the pilot study, participants were briefed and given the documentation to complete, and were then asked to undertake the set of seven tasks they had been given on the UEL-L website. They were instructed to verbalise their thoughts according to the TA condition to which they had been allocated. Participants were observed as they were in formal testing, and all measurements were collected and recorded for all reporting methods. Following the test, participants were given the post-test questionnaires and were asked about their experiences of the test, including any suggested improvements to the test procedure or its instruments. At the end, they received the promised compensation. After each pilot test, the study documents and procedure were reviewed and modified to avoid the problems that were encountered in the previous pilot.

Feedback from participants was taken into consideration to improve the main study procedure and instruments. The following sub-sections outline and discuss the changes made to the procedure and documents of the main study as a result of this feedback.

Procedural issues

Originally, the task instruction sheet was used as a reference when a participant had a task-related query. Piloting emphasised the need to read out these instructions before the commencement of testing. Operating on the assumption that participants understood the testing process took more time than explicitly explaining the process before testing began. For example, explaining that participants should alert the evaluator before starting each task at the beginning of testing rather than reminding them to do so at the start of each individual task reduced repetition and saved time.

Materials and Equipment

The pilot study highlighted the need for a better-quality audio recording tool. During the pilot study, the laptop's internal speaker was the default sound recording tool used to record verbalisations. However, when the recordings were played back, the sound quality was found to be very poor. To overcome this issue, it was decided that later experiments would use an external microphone, as the quality of these recordings would be considerably better and the external microphone could be placed closer to the participant.

Two participants suggested that providing a mouse during the session would improve the experiment. One participant stated that he 'did not like using a laptop touch-pad', and that

the option of a mouse beside the laptop touch-pad would be more convenient. Consequently, the researcher decided to provide a mouse and let the participants choose between the touch-pad and the mouse.

Other aspects of the pilot tests went smoothly and remained part of the formal test procedure. The subsequent sections explore the findings obtained from the main study.

4.9 Results

This section addresses the results obtained from the three classic TA methods used in the study and discusses the following issues: the profiles of the study participants (subsection 4.9.1), participants' task performance (subsection 4.9.2), participants' testing experiences (subsection 4.9.3), the usability problems detected (subsection 4.9.4), the cost of employing each method (subsection 4.9.5), the relationships between sample size and the problems discovered by each method (subsection 4.9.6), and finally a correlation analysis of the usability measures employed (subsection 4.9.7). The results of this study are compared to the results of other empirical studies in the section 4.10 towards the end of the chapter.

4.9.1 Participants' Profiles

Table 4.7 summarises the demographic profiles and various traits of the participants in the three TA groups. As illustrated, 32 men (53%) and 28 women (47%) participated in the study, all of whom were students at UEA; an attempt was made to represent both genders fairly equally and to ensure the gender balance of each group. 50% of the participants were male and 50% were female in the RTA condition; these percentages were 55% and 43% in the CTA and HB conditions respectively. All participants were native English speakers. The majority (93.33%) were British, while the rest were originally from Australia (5%) and Singapore (1.67%). Participants were aged between 18 and 39, with 80% between 18 and 29 years old, and 20% between 30 and 39.

All the selected participants used the Internet on a daily basis and had done so for more than five years. Nearly all of them had worked with a library website before; 85% of them had previously used their university's library website, but none of them had ever used the

evaluated website or participated in a TA usability test before. Due to having experience with the type of site used as the test object (a university library website) and being part of the target group (university students), but being novice users of the targeted website, the participants were suitable for testing the usability of the UEL-L website.

Table 4.7: Summary statistics of demographic characteristics of participants

Characteristics		**CTA** (n=20)	**RTA** (n=20)	**HB** (n=20)	**Total** (n=60)	**Percent**
Country	Britain	18	20	18	56	93.33
	Australia	1	0	2	3	5.00
	Singapore	1	0	0	1	1.67
Gender	Male	11	10	11	32	53.33
	Female	9	10	9	28	46.66
Age	18-29	15	18	13	48	80
	30-39	5	2	7	12	20
Internet use	Daily	20	20	20	60	100

The researcher believes that the independent groups were matched successfully, given that a non-parametric Kruskal-Wallis H test (Kruskal and Wallis, 1952) with an alpha level of 0.05 (i.e., $p < 0.05$)[16] revealed no statistically significant difference between the TA groups in terms of nationality ($\chi^2(2)= 2.10$, $p= .34$), gender ($\chi^2(2)= .13$, $p= .93$), age ($\chi^2(2)= 3.48$, $p= .17$), and or Internet use ($\chi^2(2)= .00$, $p= 1.0$). Therefore, the internal validity of the study is high.

4.9.2 Task Performance

As mentioned in Section 1.5, task performance measures are often used to assess reactivity associated with TA methods, which refers to a change in participants' task performance affected by the double workload of having to perform tasks and TA simultaneously (Fox et al., 2011). To measure task performance, the number of successful task completions and the time spent on tasks in this study were collected. The RTA participants in the silent

16 Most usability peer-reviewed journals typically suggest an alpha level of .05 (Sauro, 2015).

condition were the control group, with results from the other two groups compared against the RTA group's results. By having the CTA and HB groups thinking aloud while performing their tasks, the issue of reactivity would be examined on two fronts. The following subsections show the results of both indicators.

4.9.2.1 Task Completion

The task completion rate, also known as the success rate, was measured based on the number of tasks successfully completed by participants in each TA group. Participants' task completion was categorised as either successful or unsuccessful. For each successfully completed task, a participant was given a score of 1, and for each failed task, a participant was given a score of 0. Table 4.8 illustrates the task completion rates for each of the three groups.

Each participant was asked to perform seven tasks on the targeted website, meaning that 140 tasks in total were performed by each group. Participants in the RTA group successfully completed 89 tasks out of 140 tasks (a 63% success rate); the CTA participants were able to successfully complete 98 tasks (a 70% success rate); and the HB participants completed 95 tasks (a 67% success rate). In other words, each participant in the RTA group completed an average of 4.45 out of the seven tasks; each participant in the CTA group completed an average of 4.90 out of seven tasks; and each participant in the HB group completed an average of 4.75 tasks (see Table 4.9). The most difficult task (Task 5), which involved locating an article using advanced search functions, was completed successfully by only 13 of the study's 60 participants. In contrast, the easiest task (Task 2), which involved navigating the site to book a study room, was completed by 42 out of the 60 participants. None of the participants fully completed all the tasks.

Table 4.8: Descriptive statistics of task completion rates for the TA methods

Task success	CTA	RTA	HB
Total number of tasks	140	140	140
# of successful tasks	98	89	95
Percent of successful tasks	70%	63%	67%

To investigate whether a statistically significant difference existed between the total number of tasks completed by the TA condition, a one-way Analysis of Variance (ANOVA)

test was conducted. The one-way ANOVA is a parametric test used to compare the means of three or more unrelated groups, and assumes the approximate normal distribution of data and the homogeneity of variances (Field, 2009).

Normality indicates a symmetrical, bell-shaped curve, which has the highest frequency of scores in the middle with smaller frequencies toward the extreme (Field, 2009). While it can be tested visually through graphical histograms or a Q-Q plot, normality can be assessed more objectively by obtaining the *p*-value of a Shapiro-Wilk test, which is particularly recommended for sample sizes of less than 50 participants (Field, 2009). For data distribution to qualify as approximately normal, the *p*-value of the Shapiro-Wilk test must be more than 0.05 (i.e., $p > 0.05$) for each group of the independent variable[17]. Assessing the homogeneity of variances assumes that the spread of the dependent scores is roughly equal in all groups of the independent variable. Levene's test can be used to examine the accuracy of this assumption with regard to groups of normally distributed data. If Levene's test yields a significant result (i.e., $p < 0.05$), it can be said that the variances are significantly different and that the assumption of homogeneity has been violated. On the other hand, if Levene's test does not yield a statistically significant result (i.e., $p > 0.05$), it can be concluded that the variances are equal and that the assumption of homogeneity has been satisfied[18] (Field, 2009).

The Shapiro-Wilk test showed that the task success rates were not approximately normally distributed for the three TA groups, with $p = 0.003$ for the CTA group, $p = 0.024$ for the RTA group, and $p = 0.009$ for the HB group, respectively. Since the task success rates data were not normally distributed, it was not appropriate to use the one-way ANOVA test. Instead, the Kruskal Wallis H test, the alternative non-parametric test to ANOVA, was carried out, which does not assume the normal distribution of the data set (Field, 2009).

The Kruskal-Wallis H-test found no statistically significant difference in the number of successful task completions between the three TA conditions, as shown in Table 4.9. Participants' rates of task completion were therefore not affected by the double workload of simultaneously having to think aloud and carrying out tasks. The CTA and HB

17 Combining the scores of the groups and testing the combined group for normality is not an adequate measure.
18 If the sample size in each group is similar, violation of this assumption tends not to be a serious issue (Filed, 2009).

participants performed their tasks as successfully as those in the RTA group. This finding lends support to Ericsson and Simon's (1993) argument that thinking aloud does not have an effect on task performance.

Table 4.9: Inferential statistics of the task completion for the TA methods

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Task success	4.90	1.34	4.45	0.94	4.75	1.22	$\chi 2(2)=2.70$, $p=.259$

4.9.2.2 Task Time

The task time metric measured the time taken by participants in each TA condition to complete all seven tasks, regardless of whether the tasks were completed successfully (Tullis and Albert, 2008). Table 4.10 compares the total time spent on all tasks by all participants, and the average time taken to perform the tasks.

Table 4.10: Descriptive statistics of time on tasks for the TA methods

Time-on-task	CTA	RTA	HB
Overall time spent on tasks (*m*)	413.40	378.00	399.00
Mean time spent on tasks (*m*)	20.67	18.90	19.95

Task time was longest for the participants in the CTA group, with a total of 413.40 minutes, and shortest for the participants in the silent condition in the RTA group, with a total of 378 minutes. The HB group's task time fell between these.

A one-way ANOVA test was conducted in order to determine if there were significant differences in the mean time spent on all tasks. The Shapiro-Wilk test showed that the task times were approximately normally distributed for the three TA groups, with $p= 0.145$ for the CTA group, $p= 0.499$ for the RTA group, and $p= 0.061$ for the HB group, respectively. The assumption of homogeneity of variances was also met ($p= 0.866$). The one-way ANOVA test found no significance difference among the three verbalization conditions, as shown in Table 4.11. It seems that the participants in the CTA and HB conditions did not work more slowly than the ones in the RTA condition as a result of having to think aloud while performing the tasks. This finding, once again, is in line with Ericsson and Simon (1993) who stated that thinking aloud does not lead to changes in problem-solving performance.

Table 4.11: Inferential statistics of task time for the TA methods

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Time on tasks (*min*)	20.67	4.07	18.90	3.76	19.95	3.50	F(2,57)=1.96, *p*= .149

The results in this section all suggest that thinking aloud while performing tasks did not affect participants' task performance, or, in other words, did not induce reactivity. The next section discusses the testing experiences of the participants.

4.9.3 Participants' Experiences

This section reports on the measurement items in the post-test questionnaires (i.e., the System Usability Scale questionnaire and the Experience with the TA Test questionnaire), which, as mentioned earlier, sought to establish how the participants in the three TA conditions felt about: (1) the usability level of the tested website, (2) how the TA method affected their work on tasks; (3) having to think aloud (concurrently and/or retrospectively); and (4) the presence of the evaluator.

4.9.3.1 Participants' Satisfaction with the Usability of the Targeted Website

In order to gauge the effect of thinking aloud on participants' perceptions of the usability of the chosen website, participants were asked to fill out the System Usability Scale (SUS) form designed by Brooke (Brooke, 1996). The SUS form is a simple questionnaire consisting of 10 questions to be answered on a 5-point Likert scale, with 1 indicating strong disagreement and 5 indicating strong agreement (Brooke, 1996), which is widely accepted across the industry as a reliable tool for measuring the usability of computing products. However, the scores for individual items are not meaningful on their own; instead, these are compiled to yield a single score representing a composite measure of the overall usability of the system being studied. Each question has a contribution score between 0 and 4. For each of the odd-numbered questions (1, 3, 5, 7 and 9) the contribution score is calculated by subtracting 1 from the participant's Likert scale rating. For each of the even-numbered items (2, 4, 6, 8 and 10), the contribution score is calculated by subtracting the participant's Likert scale rating from 5. The sum of the contribution scores is then multiplied by 2.5 to obtain the overall SUS score. SUS scores have a range of 0 to 100, with a higher score reflecting greater participant satisfaction with a site (Brooke, 1996).

For this experiment, the standard SUS questionnaire was slightly modified by replacing the term "system" with "website" (e.g. 'I thought the website was easy to use'). To automatically calculate the SUS score for the study's multiple participants, Thomas's (2015) spreadsheet was used.

The Shapiro-Wilk test showed an approximately normal distribution for the SUS scores among the three TA groups, with p= 0.962 for the CTA group, p= 0.131 for the RTA group (silent condition), and p= 0.778 for the HB group, respectively. The assumption of homogeneity of variances was also met (p= 0.657). A one-way ANOVA test indicated that the mean satisfaction scores did not differ between the conditions (see Table 4.12). Apparently, thinking aloud while performing tasks had no effect on participants' satisfaction with the evaluated website. However, the three participant groups did not find the system very usable. The overall average SUS score of the test website was 66.20, which is under the average SUS score of 68 and indicates that the website requires improvement (Thomas, 2015).

Table 4.12: Participants' satisfaction with the tested website

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
SUS score	70.60	14.73	65.47	17.82	62.55	13.37	F(2,57)=1.39, p=.257

On a totaled scale of 1 to 100

4.9.3.2 Participant Experience with the TA Test

Since the data from the Experience with the TA Test questionnaire were not normally distributed, as revealed by Shapiro-Wilk testing (see Appendix C22), a thorough Kruskal-Wallis H-test was conducted to find out if the participant's responses differ significantly between the groups with regard to their testing experience. Table 4.13 and 4.14 present the results of participants' ratings in the three TA conditions. Note that CTA-HB and RTA-HB in Table 4.13 refer to the HB participants in the concurrent and retrospective phases of the HB condition.

To begin with, all participants were asked to assess how their working procedure on test tasks differed from their usual work approaches by estimating how much slower and how much more focused they were while working on the tasks. As shown in Table 4.14,

participants in all three conditions felt that their work on tasks was not that different from their normal work: the scores for the two items are fairly neutral, ranking around the middle of the scale, and no significant differences were found between the conditions.

Participants were next asked about the degree to which they felt having to think aloud (concurrently or/and retrospectively) was difficult, unpleasant, tiring, unnatural, and time-consuming. As shown in table 4.13, a Kruskal-Wallis H-test revealed significant differences between the conditions for "time-consuming". Accordingly, the researcher performed pairwise comparisons using Dunn's method (1964) with a Bonferroni correction in order to determine which differences conditions were significant. This post hoc analysis indicated that the participants in the RTA-HB phase found thinking aloud retrospectively to be more time-consuming than did participants in the CTA-HB phase and participants in the CTA and RTA conditions ($p< 0.05$). This difference may be explained by the longer duration of the HB test and the request for participants to provide dual elicitations, which may have caused the HB participants to rate the TA experience in the retrospective phase as more time-consuming than in the concurrent phase, and as more time-consuming than did participants in the other two conditions. For other items, the participants rated their experiences with thinking aloud as neutral to positive on average. This meant that participants in the CTA and the CTA-HB conditions did not experience reactivity while carrying out tasks.

Table 4.13: Participants and the TA test experience

	CTA		**RTA**		**CTA-HB**		**RTA-HB**		**Value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Think-aloud experience									
Difficult	2.60	0.88	2.35	1.26	2.50	1.19	2.20	1.32	$\chi2(3)=3.63$, $p=.304$
Unnatural	3.05	0.94	2.75	0.85	3.30	0.80	2.90	1.16	$\chi2(3)=4.33$, $p=.228$
Unpleasant	2.65	1.38	2.40	1.56	2.45	1.14	3.00	1.37	$\chi2(3)=2.91$, $p=.406$
Tiring	2.50	1.19	2.00	0.85	2.30	0.97	2.80	1.36	$\chi2(3)=3.81$, $p=.282$
Time-consuming*	2.70	1.48	3.05	1.30	2.90	1.43	4.25	0.91	$\chi2(3)=11.36$, $p=.010$

Five-points scale (1: Strongly disagree to 5: Strongly agree) * $p< 0.05$ significance obtained

The final part of the Experience with the TA Test questionnaire included measurement items about the presence of the evaluator. Participants were asked to indicate to what extent they found it unpleasant, unnatural and disturbing to have the evaluator present during the study. Kruskal-Wallis H-test testing yielded no significant differences between the

conditions regarding these questions. As the average scores of the participants ranged between 1.10 and 1.80, the participants clearly felt that the evaluator's presence did not affect their testing experience.

Table 4.14: Participants' experience with the TA test

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Working condition on tasks							
Slower than my normal working	2.40	1.09	2.15	1.30	2.65	1.22	$\chi 2(2)=2.66$, $p=.264$
More focused than my normal working	3.05	1.14	2.80	1.36	3.20	1.70	$\chi 2(2)=8.98$, $p=.638$
Evaluator presence							
Unnatural	1.35	0.81	1.80	1.21	1.50	0.88	$\chi 2(2)=2.51$, $p=.302$
Disturbing	1.20	0.44	1.60	0.50	1.40	0.51	$\chi 2(2)=1.94$, $p=.378$
Unpleasant	1.10	0.30	1.30	0.57	1.25	0.44	$\chi 2(2)=1.90$, $p=.386$

To summarise, participants in all three usability testing conditions reported similar results across testing conditions. Most measures included in the questionnaire yielded neutral to positive responses for all three conditions. The only significant difference found was that the HB participants felt that thinking aloud retrospectively was more time-consuming than did participants in other conditions. The next section will discuss the usability problems identified by each TA method.

4.9.4 Usability Problems

This section presents the results relating to the quantity and quality of usability problem data at the level of individual problems (i.e., problems detected per participant in each condition) and final problems (i.e., the aggregate problems detected in each condition). Usability problems were determined using the process described in section 3.10. Five different indicators were used to evaluate the usability problems discovered by the three testing methods: 1) the number of problems, 2) the sources of problems, 3) the severity of problems, 4) the type of problems, and 5) the uniqueness of problems. Since Shapiro-Wilk testing revealed that the individual usability problem data were not normally distributed (see Appendix C23), a Kruskal-Wallis H-test was used to analyse the data. Descriptive data is presented for the final problem set.

4.9.4.1 Individual Usability Problems

The most common way to measure usability issues is to count the number of problems found (Tullis and Albert, 2008). Table 4.15 presents the mean number and standard deviation for problems detected per participant, and classifies all problems according to how they were detected: (1) through observation (i.e., from observed evidence with no accompanying verbal data), (2) through verbalization (i.e., from verbal data with no accompanying behavioural evidence), or 3) through a combination of observation and verbalization.

As can be seen in table 4.15, a Kruskal-Wallis H-test revealed significant differences in the number of individual problems detected by participants between testing approaches. Accordingly, the researcher performed pairwise comparisons using Dunn's (1964) method with a Bonferroni correction in order to determine which differences were significant. This post hoc analysis indicated that the RTA participants discovered significantly fewer individual problems than participants in the CTA and HB conditions ($p < 0.05$). A possible explanation for this discrepancy is that asking test participants to report problems after performing tasks silently may have increased their likelihood of forgetting to report problems during the retrospective phase, even if they had noticed these problems while performing tasks. This finding lends support to Ericsson and Simon's (1993) argument that vital information may be lost when applying retrospective research methods, and casts doubt on the validity of the outcome of a RTA evaluation as an overall indication of usability.

However, no significant differences were detected between the results of the HB and CTA conditions, suggesting that thinking both concurrently and retrospectively did not cause the HB participants to detect a substantially larger or smaller number of individual problems than the CTA participants. The HB participants not finding a significantly larger number of individual problems may be attributed to their feeling that they had already provided detailed comments in the concurrent phase, and/or feeling tired due to the prolonged duration. The fact that the HB participants did not detect a significantly smaller number of problems than the CTA participants could be attributed to their providing a full account during the concurrent reporting phase, which led them to detect a comparable number of problems to the CTA participants. These explanations are supported by the final

problems produced by the concurrent and retrospective phases in the HB condition, as presented in section 4.9.4.2.

Table 4.15: TA methods and the number of individual problems

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Observed	1.35	0.74	1.30	0.47	1.20	0.41	$\chi2(2)= 5.15$, $p= .773$
Verbalised*	2.65	1.75	1.00	1.25	2.75	2.48	$\chi2(2)= 10.08$, $p= .004$
Both	5.55	1.63	4.05	1.98	5.95	3.82	$\chi2(2)= 5.30$, $p= .071$
Total*	9.55	3.26	6.35	3.09	9.90	5.33	$\chi2(2)= 8.21$, $p= .016$

* p< 0.05 significance obtained

With respect to the manner in which the individual problems were detected, it can be seen from table 4.15 that participants' verbalisations in all three conditions aided them in detecting problems that were not otherwise observed (verbalised problems), or in emphasising or explaining problems that were also observed in their actions (combined problems). This result confirmed the invaluable contribution of verbal protocols to the outcome of usability testing that numerous scholars have highlighted in previous research (e.g., Nielsen, 1993; Rubin, 1994; Dumas and Redish, 1999; Barnum, 2002).

A Kruskal-Wallis H-test revealed highly significant differences between one or more of the conditions regarding the number of individual problems detected, while a Bonferroni post hoc analysis showed that the CTA and HB participants detected a significantly higher number of verbalised individual problems than the RTA participants ($p< 0.05$). There were no differences in the number of individual problems detected through evaluator observation or the combined source. However, as the CTA and HB participants did not experience more observable difficulties than the RTA participants, this once again supports Ericsson and Simon's (1993) argument that thinking aloud while performing tasks does not negatively affect performance.

Individual usability problems and severity levels

The severity levels of individual problems were categorised into one of four types according to their impact on participants' performance: 1) critical, 2) major, 3) minor, and 4) enhancement (Dumas and Redish, 1999; Zhao et al., 2012), as outlined in Table 4.16.

Table 4.16: Coding scheme for problem severity levels

	Problem Severity level	Definition
1	**Critical**	The usability problem prevented the completion of a task
2	**Major**	The usability problem caused significant delay (more than one minute) or frustration
3	**Minor**	The usability problem had minor effect on usability, several seconds of delay and slight frustration
4	**Enhancement**	Participants made suggestions or indicated a preference, but the issue did not cause impact on performance

When assigning severity levels to individual problems, the persistence of each problem, which refers to the number of times the same problem is encountered by a test participant, was also taken into consideration (Hertzum, 2006). For example, if the same participant encountered the same problem more than three times, even if each incident only had a minor impact, the individual problem was considered as major due to the aggregation of impact (Nielsen, 1993). Table 4.17 presents the mean value and the standard deviation of the number of individual problems at each severity level. A Kruskal-Wallis H-test and a post hoc analysis showed that the CTA and HB participants found a significantly higher number of minor problems than the RTA participants ($p< 0.05$). There were no significant differences between the methods for the number of individual critical, major or enhancement problems detected.

Table 4.17: TA methods and individual problem severity levels

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Critical	1.90	0.74	2.20	0.83	2.15	0.91	$\chi2(2)= 1.96$, $p= .375$
Major	2.90	1.74	2.15	1.84	2.50	2.55	$\chi2(2)= 2.31$, $p= .314$
Minor*	4.40	3.74	1.80	1.63	4.65	4.30	$\chi2(2)= 8.55$, $p= .014$
Enhancement	0.35	0.48	0.20	0.62	0.60	1.48	$\chi2(2)= 0.90$ $p= . 933$

* $p< 0.05$ significance obtained

Individual usability problem types

Two independent usability experts were asked to classify the detected problems from the study into four types, as outlined in table 4.18. These types are based on an initial review of the data, the literature related to the categorisation of usability problem of online libraries (Van den Haak et al., 2004), and the literature related to the categorisation of website usability problems (Tullis and Albert, 2008; Zhao et al., 2012). The experts were

informed that when they thought that a problem should be coded into a new category, they should feel free to do so.

Table 4.18: Problem types coding scheme

Problem type	Definition	Example
Navigation	Participants have problems navigating between pages or identifying suitable links for information/functions.	The participant has trouble returning to the home page
Layout	Participants encounter difficulties due to web elements, display problems, visibility issues, inconsistency, and problematic structure and form design	The participant feels that the font is too small
Content	Participants think certain information is unnecessary or is absent; Participants have problems understanding the information including terminology and dialogue	The participant does not understand the feedback of an error messages
Functionality	Participants encounter difficulties due to the absence of certain functions or the presence of problematic functions	The participant expects an option on 'Catalogue' page to specify how many items to load per page

Inter-coder reliability was computed using Cohen's kappa, a reliability measure based on the assumption that each coder is classifying the same problem or that the total number of problems that need to be coded is known or can reliably be estimated (Barendregt et al., 2006). SPSS was used to calculate the agreement value (Robson, 2002). Robson (2002) outlines the agreement levels of kappa values as:

(1) < 0.40: poor agreement

(2) 0.40-0.60: fair agreement

(3) 0.60-0.75: good agreement

(4) > 0.75: excellent agreement

The overall kappa value was 0.87, which shows a highly satisfactory level of inter-coder agreement. The coders discussed the problems that were classified in different categories and created a final classification of all problems on which they both agreed.

Table 4.19 shows the number of different types of individual problems identified in the classical TA methods. In all conditions, navigation clearly presented the most problems to the participants. This is likely because in working with the tested site, the participants had to navigate many menus of links, each of which they had to interpret before being able to

move on to the next level. A Kruskal-Wallis H-test and Bonferroni post hoc analysis showed significant differences between the conditions regarding layout problems: both the CTA and HB participants reported more layout problems than participants in the RTA condition ($p< 0.005$), with the verbalisation conditions bringing to light the other three problem types with similar frequency.

Table 4.19: TA methods and individual problem type

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Navigation	4.55	3.42	3.85	3.34	4.90	3.56	$\chi2(2)=0.99$, $p=.607$
Layout*	3.10	2.22	1.00	0.85	3.25	3.20	$\chi2(2)=12.55$, $p=.002$
Content	0.85	0.48	0.60	0.59	0.55	0.60	$\chi2(2)=3.612$, $p=.164$
Functionality	1.05	0.82	0.90	0.44	1.20	1.32	$\chi2(2)=0.45$, $p=.795$

* $p< 0.005$ significance obtained

4.9.1.2 Final Usability Problems

After analysing all of the usability problems found across conditions, the number of problems encountered by all participants were collected, excluding any repeated problems to arrive at a total number of final usability problems. In total, 75 final usability problems were extracted from the test sessions in the three TA conditions. Participants in the CTA condition identified 47 out of the 75 final problems (62%), 13 of which were unique problems, which were found only by the CTA participants. Participants in the RTA condition identified 33 final problems (44%), 8 of which were unique problems, while participants in the HB condition identified 52 final problems, 17 of which were unique problems (see Table 4.20). Therefore, with respect to the detection of final problems, the CTA and HB methods were again more successful than the RTA method. As the CTA and HB methods only differed by 5 final problems, it is fair to say that these two methods revealed a similar number of final problems in the UEL-L site.

Further analysis of the HB condition results revealed that 25 of the 52 total final problems (48%) were detected in the concurrent phase, whereas 5 problems (10%) were only found in the retrospective phase, and 22 problems (42%) were duplicated between both phases, meaning that the majority of the final problems (90%) were in fact detected in the concurrent phase. This reinforces the claim that the retrospective phase has a limited capacity to contribute to usability problem detection, and that the combination of

concurrent and retrospective phases advised by Ericsson and Simon (1993) may be less beneficial than expected in terms of the quantity of usability problems detected.

Table 4.20: TA methods and the number of final problems

	# of problems	% of problems	# of unique problems	% of unique Problems
CTA	47	62 %	13	17 %
RTA	33	44 %	8	10 %
HB	52	69 %	17	22 %
Total	75	100 %	38	50 %

Although there were 20 problems (26%) that occurred in all of the three conditions, the overlap between two rather than three conditions was considerably less, ranging from 2% to 16%. These low percentages indicate a substantial number of unique problems identified by three conditions (38 problems). This result is perhaps not very surprising given the quality and quantity of pages on the tested website. The HB participants discovered twice as many unique problems as the RTA participants. The Venn diagram in Figure 4.7 shows the overlap between the three conditions. Appendix C20 lists the final problems discovered by the participants in this study.

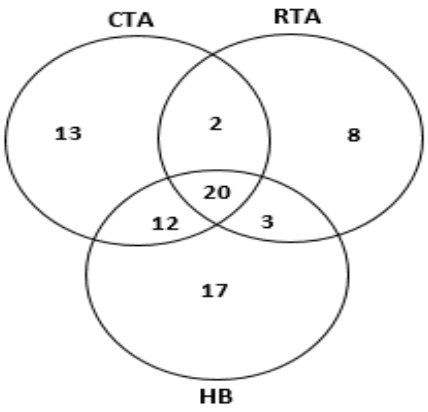


Figure 4.7: Venn diagram showing overlap in problems between think-aloud protocols

Final usability problems and their sources

Final usability problems were coded according to verbalisation source, observation source, and a combination of both. A problem was deemed to have a combined source if the individual problems had been emerged from both verbal and observation sources. To qualify as having either a verbal or observed source, a final problem had to consist of

individual problems from a single source of origin (all verbal or all observed) (Zhao et al., 2012). Table 4.21 sets out the final problem sources and their definitions.

Table 4.21: Final problem sources coding scheme (Zhao et al., 2012)

Final problem source	Definition
Observation	All component individual problems of a final problem were from the observation source only
Verbalisation	All component individual problems of a final problem were from the verbalisation source only
Combination of both	Component individual problems of a final problem were from a mixture of verbalisation source and observation source

As shown in Table 4.22, the results for the CTA condition were that 6 problems were derived from observation evidence, 15 from verbal evidence and 26 from a combination of the two. In the RTA condition, 7 problems were derived from observation evidence, 6 from verbal evidence and 20 from a combination of the two. In the HB condition, 3 problem were derived from observation evidence, 17 from verbal evidence and 32 from a combination of the two. While the CTA (15 problems) and HB (17 problems) encouraged more verbalised final problems than the RTA (6 problems), a larger number of the unique problems in the CTA (69%), the RTA (62%), and the HB (82%) conditions were derived from verbalisation. With respect to the 5 problems detected in the retrospective phase in the HB condition, all of these were derived from verbalisation.

Table 4.22: TA methods and final problem sources

	CTA		RTA		HB	
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**
Observed	0	6	0	7	0	3
Verbalised	9	6	5	1	14	3
Both	4	22	3	17	3	29
Total	13	34	8	25	17	35

Final usability problems and severity levels

The assignment of severity levels to final problems must take into account the discrepancies between how a given problem may be experienced by participants; for example, one participant may circumvent a problem very quickly, while another may spend a long time overcoming the same problem. To bypass potential conflict between severity levels, levels were assigned according to the majority (Lindgaard and

Chattratichart, 2007). In those cases where the contradictory severity levels emerged with an equal number of participants, assignment took place according to the highest severity level (Ebling and John, 2000).

Table 4.23 presents the number of problems according to severity level for the three TA conditions. As shown in the table, while the three methods identified similar numbers of critical problems, the distribution of severity differed between each method. 28% (13 problems) of the final problems from the CTA method were high impact problems (with critical and major effects), and 70% (34 problems) were low impact problems (with minor and enhancement effects). For the RTA condition, 39% (13 problems) of final problems were high impact, and for the HB condition, 23% (12 problems) of final problems were high impact. The final five problems found only in the retrospective phase in the HB condition were all minor problems.

Table 4.23: TA methods and final problem severity levels

	CTA		**RTA**		**HB**	
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**
Critical	0	2	0	2	0	2
Major	2	9	2	9	3	7
Minor	9	21	5	13	12	23
Enhancement	2	2	1	1	2	3
Total	13	34	8	25	17	35

Regarding unique problems, analysis indicated that no one method identified critical problems that were not identified by the other methods. Analysis also revealed that 15% of the unique problems identified by CTA participants were high impact problems, 25% of the unique problems identified by RTA participants were high impact, and 17% of the unique problems identified by HB participants were high impact.

With respect to the sources of unique problems, 69% (9 problems) of those found by CTA participants were derived from verbalisation, with 88% (8 problems) of these being low impact. 62% (5 problems) of unique problems found by RTA participants and 82% (14 problems) of those found by HB participants were derived from verbalisation and were all low impact (see Table 4.24).

Table 4.24: Sources and severity levels for the unique final problems in the three TA conditions

	CTA			RTA			HB		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Critical	0	0	0	0	0	0	0	0	0
Major	0	1	1	0	0	2	0	0	3
Minor	0	6	3	0	4	1	0	12	0
Enhancement	0	2	0	0	1	0	0	2	0
Total	0	9	4	0	5	3	0	14	3

Final usability problem types

Table 4.25 shows the number of final usability problems for each problem type according to each TA condition. Of the 75 final problems detected, there were 20 navigational problems, 28 layout problems, 14 content problems, and 13 functional problems. CTA and HB participants identified more problems of each type than RTA participants. The distributions of problem types were similar in the CTA and RTA conditions, with the least frequent being content, then functionality, then layout, and finally navigational problems being the most frequent. The HB condition showed a similar pattern, with the exception of layout problems being the most frequent and navigational problems being the second most frequent. In terms of the unique problems found by the three methods, HB participants seemed to detect more unique layout problems than CTA and RTA participants. With regard to the problems generated from the retrospective phase of the HB condition, three of these were layout problems and two were content problems.

Table 4.25: TA methods and final problem types

	CTA		RTA		HB		**Total**
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**	
Navigation	3	14	2	10	1	14	20
Layout	5	10	2	7	8	12	28
Content	3	3	3	2	5	2	14
Functionality	2	7	1	6	3	7	13
Total	13	34	8	25	17	35	75

Figures 4.8, 4.9 and 4.10 depict the final problems detected according to their types and severity levels in each TA method. As shown here, all of the critical problems found by the three methods related to navigation.

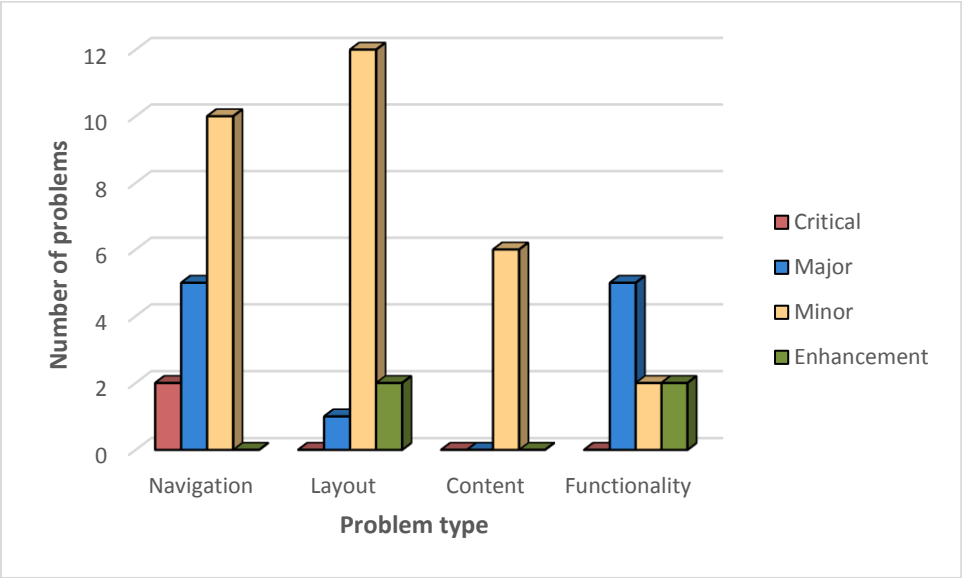


Figure 4.8: Types and severity levels for the final problems in CTA condition

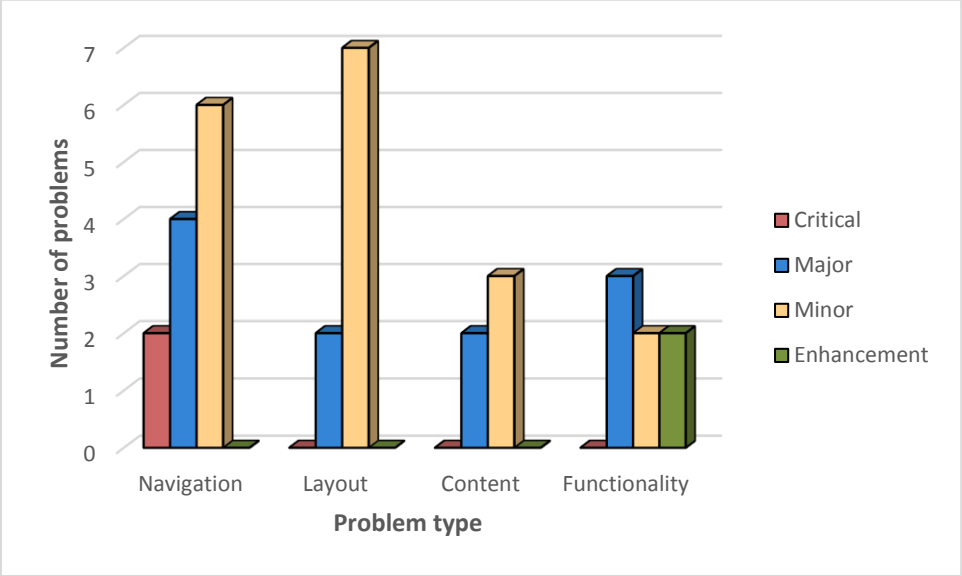


Figure 4.9: Types and severity levels for the final problems in RTA condition

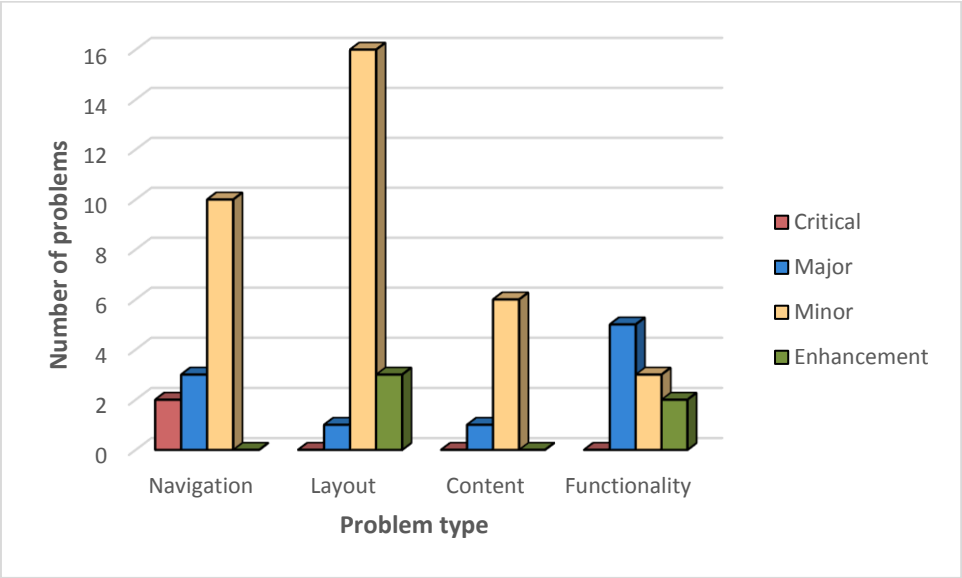


Figure 4.10: Types and severity levels for the final problems in HB condition

Table 4.26 provides a breakdown of the unique problems (38 problems) according to their problem sources and types, and shows that all unique navigational problems were derived from verbalisation.

Table 4.26: Sources and types for the unique final problems in the three TA conditions

	CTA			RTA			HB		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Navigation	0	3	0	0	2	0	0	1	0
Layout	0	2	3	0	2	0	0	7	1
Content	0	3	0	0	1	2	0	3	2
Functionality	0	1	1	0	0	1	0	3	0
Total	0	9	4	0	5	3	0	14	3

Further analysis of the types and severity levels of unique problems indicated that for the CTA and RTA conditions, all problems relating to layout, content and functionality were at low severity levels, as shown in table 4.27.

Table 4.27: Types and severity levels for the unique final problems in the TA conditions

	CTA				RTA				HB			
	Critical	**Major**	**Minor**	**En.***	**Critical**	**Major**	**Minor**	**En.**	**Critical**	**Major**	**Minor**	**En.**
Navigation	0	2	1	0	0	2	0	0	0	0	1	0
Layout	0	0	3	2	0	0	2	0	0	1	6	1
Content	0	0	3	0	0	0	3	0	0	1	4	0
Functionality	0	0	2	0	0	0	0	1	0	1	1	1
Total	0	2	9	2	0	2	5	1	0	3	12	2

*Enhancement

Figure 4.11 illustrates some of the problems identified by the participants on the evaluated website. A report on these problems was sent to the website administrator, who then sent an appreciation letter in response (see Appendix C21).

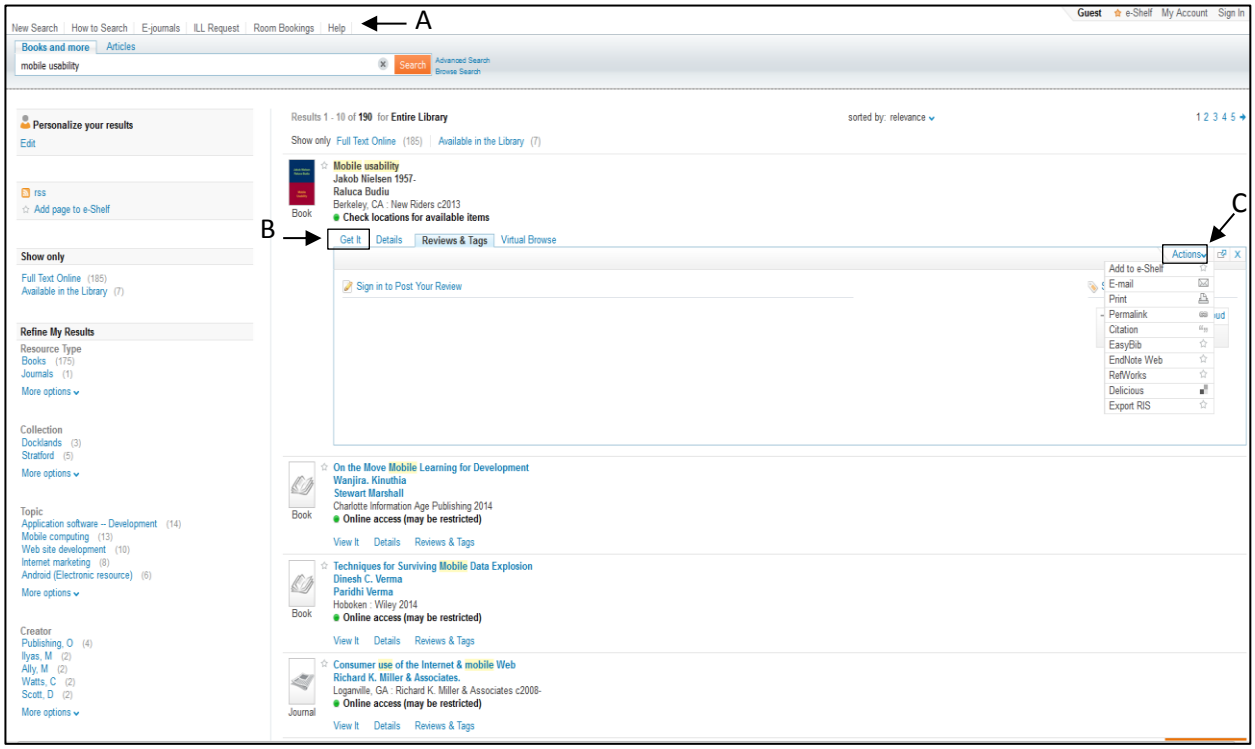


Figure 4.11: Illustration of some usability problems discovered: A) No 'Home' page tab; B) the link 'Get it' is problematic because users thought that by clicking on this link they could view an electronic copy of an item; C) the link 'Action' is problematic because many users failed to click on it to find information about item citations.

Reliability of problem identification and classification

As mentioned in section 3.10, an extra evaluator was recruited to carry out an inter-coder reliability check on usability problem analysis. Nielsen (1992) indicates that while there is no official certification for usability specialists, people with graduate degrees who have several years of work experience in the usability area' can be classified as such. Nielsen also found that usability specialists are better at finding usability problems than people without such expertise, and that double specialists, who have experience both of usability and the interface being investigated, perform even better in this respect.

The independent evaluator in this study was a PhD student in the area of usability testing under the supervision of Dr Pam Mayhew. He has several years of usability experience

and is well informed of the literature on usability evaluation, and had published a number of scientific articles prior to the commencement of this research. As recruiting experts in usability evaluation who are also familiar with the interface under study can be difficult (Stone et al., 2005), the researcher asked the independent evaluator to familiarise himself sufficiently with the tested system in advance of his evaluation.

The researcher introduced the independent evaluator, who was employed on a voluntary basis, to the use of the problem analysis approach (see Section 3.10). Following this, a guide to the approach was sent out to him in MS Word format. The guide consisted of five sections:

1. Overview of the usability problem analysis procedure;
2. Usability problem definition, problem indicator lists, problem matching criteria, and the coding schemes for problem sources and severity levels;
3. Instructions for playing the video recordings of testing sessions;
4. Problem report templates, with instructions for how to write problem reports; and
5. The task descriptions list and the steps for optimal performance.

The independent evaluator borrowed a laptop in which the tested data was installed, independently coded the usability problems for the first participant, and discussed his disagreements with the researcher. Subsequently, the independent evaluator analysed six randomly selected testing videos (two from each condition). The minimum reliability check sampling is recommended to be at least 10% of the full sample size (Lombard, 2004), to which this study adhered. On completion, the author and the independent evaluator compared their individual sets. The any-two agreement formula provided by Hertzum and Jacobsen (2001) was used to calculate inter-coder reliability across the six videos:

$$Any - two\ agreement = \frac{|Pi \cap Pj|}{|Pi \cup Pj|} \quad (1)$$

In this equation, Pi and Pj are the problems identified by evaluators "i" and "j" respectively. Its value ranges from 0% in the case of no agreement amongst the evaluators to 100% in the case of full agreement.

The evaluators then independently grouped the individual problems based on the matching criteria to form final problems. Upon the completion of this step, another meeting was held to compare the final problem sets of the two coders, and the any-two agreement for final problem production was calculated.

The average any-two agreement for individual problem identification across the six videos was 67% (individual agreements were 70%, 63%, 69%, 74%, 66%, and 58%). The any-two agreement for final usability problem production was 72% (CTA: 70%, RTA: 78%, and HB: 68%). Overall, the agreements are high compared to those set out in Hertzum and Jacobsen's (2001) study, wherein agreements between evaluators ranged from 5% to 65%. The reliability of the coding of the problem source and severity level was examined using Cohen's kappa. For individual problems, the kappa value for problem sources was 0.819, and 0.654 for problem severity. For final problems, the kappa value for problem sources was 0.826, and 0.693 for severity. These values reveal a high degree of reliability for the coding.

4.9.5 Comparative Cost

The cost of employing the three TA methods under study was measured by recording the time the evaluator spent conducting testing and analysing the results for each method. As mentioned in section 3.9.4, session time, recorded via an observation sheet (Appendix C18), refers to the time required to carry out full testing sessions, including the instruction of participants, data collection, and solving any problems that may arise during the session. Analysis time, collected via web-based free time tracking software called "Toggle" [19](Version 2013), refers to the time required to extract usability problems from each method's testing data. The collected data from these measures were then utilised for a comprehensive evaluation of the financial costs of each methods. The following subsections review the approximate time required for each TA method (section 4.9.5.1) and estimate their financial costs (section 4.9.5.2).

[19] https://toggl.com/

4.9.5.1 Temporal Cost

Table 4.28 shows the time spent by the researcher on applying and analysing the results for the three verbalisation methods. As is clear from the table, the CTA method required the shortest session time (640 minutes), whereas the HB method required the longest session time (1233 minutes). The RTA testing lasted for 1164 minutes. The mean times of RTA sessions (58 minutes) and HB sessions (61 minutes) were almost double that of CTA sessions (32 minutes) (see Table 4.29). The total time taken to apply the three methods was 3037 minutes.

Table 4.28: TA methods and time expense

	CTA	**RTA**	**HB**	**Total**
Session time *(m)*	640	1164	1233	3037
Analysis time *(m)*	733	1081	1150	2964
Total time *(m)*	1373	2245	2383	6001

An ANOVA test was conducted to determine if there were significant differences in the mean session time between the the conditions. A Shapiro-Wilk test showed that the data were approximately normally distributed for the three conditions, with $p = .223$ for the CTA condition, $p = .470$ for the RTA condition, and $p = .523$ for the HB condition, respectively. The assumption of homogeneity of variances was also met ($p = .439$). As expected, the session time significantly differed between the three groups. To examine which of these groups differed from each other, a Tukey post hoc analysis was conducted, revealing that RTA and HB session times were significantly longer than CTA session times. No significant difference was found between the RTA and HB conditions. Given that the analysis of task time revealed no significant difference between the conditions, as mentioned in section 4.9.2, this discrepancy can be attributed to the fact that RTA and HB participants had to watch a video recording of their performance in order to provide retrospective reporting, which obviously prolonged the session times.

Table 4.29: Session time for the TA methods

	CTA		**RTA**		**HB**		**Value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Session time (*m*)*	32.00	6.00	58.20	7.57	61.65	8.64	$F(2,57)=93$, $p<0.0001$

* $p< 0.0001$ significance obtained

Identifying usability problems through analysis of the videos was the most time-consuming activity in this study. The video footage of the evaluation sessions came to a total of more than 1997.4 minutes, being 413.4 minutes of recordings of evaluations by CTA participants, 780 minutes by RTA participants and 804 minutes by HB participants. The total time taken to identify usability problems using the three methods was 2964 minutes, with the HB method requiring the most time (1150 minutes) in comparison to the CTA (733 minutes) and RTA methods (1081 minutes). Once assumptions of the normality and homogeneity of variance were met, ANOVA testing and a Tukey post hoc analysis were conducted, concluding that analysis time was significantly longer for the HB condition than for the CTA and RTA conditions. The longer analysis times for the RTA and HB conditions is not surprising, as prolonged session times will inevitably lead to a longer analysis process and the evaluator had to thoroughly review each testing video in order to detect usability problems.

Table 4.30: Analysis time for the TA methods

	CTA		RTA		HB		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Analysis time (m)*	36.65	4.91	54.05	3.36	57.50	4.83	$F(2,57)=49$, $p<0.0001$

* $p< 0.0001$ significance obtained

The overall results showed that the CTA method incurred the shortest time (1373 minutes), followed by the RTA method (2245 minutes) and then the HB method (2383 minutes). The total time taken for testing and analysis of the three methods was 6001 minutes. Time per problem can be calculated by dividing the time the evaluator spent on a method by the number of problems identified by that method (Als et al., 2005). The CTA method required 29 minutes per usability problem, whereas the RTA method required 68 minutes per usability problem and the HB method required 45 minutes per usability problem (see Table 4.31). Therefore, based on the results presented, the outcomes and the time and effort required by the evaluator favour CTA testing over RTA and HB testing.

Table 4.31: TA methods' temporal costs per problem

	Time spent (*m*)	**Problem found**	**Time per Problem (*m*)**
CTA	1373	47	29
RTA	2245	33	68
HB	2383	52	45
All	6001	75	80

4.9.5.2 Financial Cost

Financial constraints define the modern business environment and dictate the extent to which a company can improve its productivity. For example, securing a slightly higher quality outcome at a much larger cost would not necessarily be considered a cost-effective alternative. Martin et al. (2014) places the daily rate charged by usability evaluators for usability consultation at £800.00 per 7.5-hour day, or approximately £107 per hour. This figure can be compared to the data from Section 4.9.5.1 to produce the financial costs for conducting these testing methods in a business environment. Table 4.7 shows the amount of evaluator hours spent conducting and analysing the results of each method multiplied by the hourly cost of a usability evaluator to produce the total financial cost of each TA evaluation (rounded to the nearest pound). It can be seen from Table 4.32 that CTA testing would cost £2448, which is substantially less than the cost of the other two methods: £4248 for the HB method and £4002 for the RTA method. The cost of the application and analysis of all three methods would be £10698.

Table 4.32: TA methods' financial cost

	Evaluator Minutes	Evaluator Hours	Hourly Fee	Financial Cost
CTA	1373	22.88	£107	£2448
RTA	2245	37.41	£107	£4002
HB	2383	39.71	£107	£4248
All	6001	100.00	£107	£10698

By comparing the financial costs of each method against the number of problems detected, the financial cost per problem can be deduced (Martin et al., 2014) (see Table 4.33). The CTA method yielded the lowest costs per problem at £52, while the RTA and HB methods yielded costs of £121 and £81 per problem respectively. From the overall results, CTA testing appears to be more cost-effective than RTA or HB testing.

Table 4.33: TA methods' finical costs per problem

	Financial Cost	Problem found	Cost per Problem
CTA	£2448	47	£52
RTA	£4002	33	£121
HB	£4248	52	£81
All	£10698	75	£142

4.9.6 Relationship between Sample Size and Number of Problems Detected

Given that one of the first decisions in planning a usability test is choosing an appropriate sample size, one of the primary objectives of this study is to investigate the relationship between sample size and the number of problems detected in TA usability testing. It has been controversially stated by other researchers that five test participants are sufficient to find 85% of usability issues (e.g., Nielsen, 2000; Nielsen, 1994a). As mentioned in section 4.9.1.2, the three TA groups reported 75 usability problems on the test website, 85% of which would equate to 64 problems. None of the groups reported this many problems, though each group used twenty test participants (see section 4.10). The "five participants" argument is therefore still highly debatable.

Nevertheless, the percentages of problems detected by five participants from each of the TA methods under investigation were compared in order to highlight any similarities or differences between the performances of the methods. In addition, the overall relationship between the sample size and the number of problems discovered in each condition was examined to determine whether or not the methods showed similar patterns. A 'good' test method, in this context, is one that can assist in finding a large proportion of usability problems using as few participants as possible (wherein the total number of usability problems is roughly estimated as the sum of the usability problems identified by each method). Although it is impossible in practice to obtain a complete set of problems with one application because of the possibility of overlooking or misidentifying usability problems (Jeffries and Miller, 1998), some intriguing findings were obtained from the comparisons.

This section is organised as follows: the first subsection (4.9.6.1) examines the number of problems discovered by the best and first five participants from each TA condition, and explores the overall relationship between the sample size and number of problems discovered in each TA condition. The second subsection (4.9.6.2) calculates the number of participants needed to find 85% of problems.

4.9.6.1 Number of Problems Discovered by the Best and First Five Participants

This subsection explores the percentage of problems detected by five participants from each TA group by assessing the first and the best groups of five participants, beginning with the best as they could have been the sole selected group members for a sample size of five. Table 4.34 shows the performance of the top five participants in each TA group. The designations T-CTA, T-RTA, and T-HB in the table refer to the top performing (T) five participants who discovered the most problems for the CTA, RTA and HB conditions respectively. The T-CTA, T-RTA, and T-HB groups discovered only 29%, 21% and 32% respectively of the total number of usability problems, which is notably less than the claim of 85%. However, the T-CTA and T-HB groups performed better than the T-RTA group, in line with the overall performance of the methods. The five top performing participants selected from the three TA conditions only detected 43% of the total number of usability problems. These results confirm that the five-participant argument in usability testing is far from settled.

Table 4.34: Top (T) five participants and number of problems discovered (absolute and percentage of total number)

Top performing five participants								**(Nielsen, 2000)**		**Maximum to be discovered**	
T-CTA		**T-RTA**		**T-HB**		**All groups**					
#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
22	29%	16	21%	24	32%	32	43%	64	85%	75	100%

Figure 4.12 shows how the 20 participants within each condition performed. In order to reduce the order effect, participants' results were selected randomly using a random number generator. As shown here, the first five participants from the CTA and HB conditions identified 24% and 21% of the final usability problems detected at the time of the evaluation respectively, and once again performed better than the first five participants in the RTA conditions, who only identified 17% of the final usability problems. Furthermore, the first ten participants (double the recommended magic number) in the CTA condition found 36% of the total number of problems, 38% of the total number of problems in the HB condition, and 32% of the total number of problems in the RTA condition. As the curves in Figure 4.12 illustrate, however, participants continued to detect new problems even after the fifteenth participant; it can also be seen that the curves of the

CTA and HB conditions were very similar. Therefore, it can be argued that the relationship between sample size and percentage of problems detected is more or less the same for these two conditions, which both differ considerably from that of the RTA condition. To find 44% of the total usability problems, the RTA participants required 17 participants, as opposed to the 11 participants required in the HB condition and the 13 participants required in the CTA condition to detect the same percentage of problems.

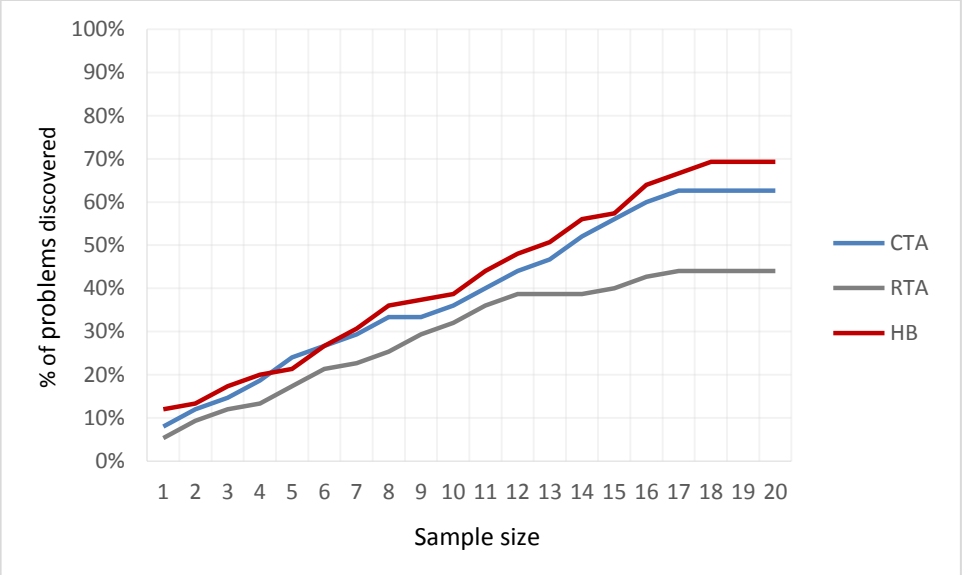


Figure 4.12: Participants' performances (cumulative) in all three conditions

The following section determines the sample size needed to find 85% of usability problems.

4.9.6.2 The Sample Size Required to Detect 85% of Problems

In order to estimate the sample size needed to detect a pre-set percentage of problems, the average detection rate of usability problems must first be calculated. This can be defined as "the average of the proportion of participants experiencing each observed problem" (Lewis, 2001, p.3). Albert and Tullis (2013, p.116) explain how the average detection rate can be calculated as follows:

> *"...line up all the usability issues discovered during the test. Then, for each participant, mark how many issues were observed [...] Add the total number of issues identified with each participant, and then divide by the total number of issues. Then, take the average for all test participants*."

Once the average detection rate is obtained, the number of participants required to detect a pre-set percentage of usability problems can be estimated using the well-known equation below, which is based on the binomial probability formula:

$$\textit{Proportion of problems to be detected (e.g., 85\%)} = 1 - (1 - P)^{n} \quad (2)$$

where *P* is the average problem detection rate and *n* is the number of participants (Turner et al., 2006). This can be calculated automatically using Sauro's online "Sample Size Calculator"[20]. The detection rate in the current study is 0.088, which means that 24 test participants would be needed from the total sample of 60 participants in order to detect 85% of the total number of usability problems found by the three groups (see Table 4.35).

Table 4.35: Participant number and the targeted percentage of problems

Targeted Parentage	Sample size required
99%	52
95%	41
90%	36
85%	24
75%	20
50%	16

An adjusted average detection rate was also calculated to estimate the sample size needed to detect 85% percentage of problems in each TA condition, as this is recommended to reduce the bias towards overestimation which occurs with small sample sizes (N ≤ 20 test participants) (Lewis, 2001, 2006b). This adjustment involves averaging based on Good-Turing discounting and a normalisation method proposed by Hertzum and Jacobsen (2001):

$$Padj = \frac{1}{2}\left[\left(Pest - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\right] + \frac{1}{2}\left(\frac{Pest}{1 + GT\ adj}\right) \quad (3)$$

where *P* adj is the adjusted *p* value, *P*est is the unadjusted P value, *n* is the sample size, and *GT* adj is the "Good-Turing" adjuster, which is the number of usability problems detected by only one participant divided by the total number of usability problems found in the study (Lewis, 2006b). Once *P* adj is calculated, it can be used again in the formula (1), where *P* is *P* adj, to estimate the number of users needed to detect a specific percentage of problems. The adjusted average detection rate of usability problems was 0.056 in the

20 http://www.measuringu.com/problem_discovery.php

CTA condition, 0.041 in the RTA condition, and 0.060 in the HB condition. In order to find 85% of the problems in the tested interface, at least 34 test participants would be needed in the CTA condition, 46 in the RTA condition, and 30 in the HB condition. That is, the RTA condition would require 12 more test participants than the CTA condition and 16 more participants than the HB condition in order to find 85% of the total number of problems (see Table 4.36).

Table 4.36: The sample size required to find 85% of the final number of problems

Targeted Parentage	CTA	RTA	HB
GT adj	0.082	0.064	0.144
P est	0.076	0.053	0.091
1/n	0.05	0.05	0.05
P adj	0.056	0.041	0.060
Sample size needed to reveal 85%	34	46	30

4.9.7 Correlational Analysis of Usability Measures

In addition to comparing the outcomes of the experimental conditions, the size of correlations between usability measures can provide further insights. This point has notably been made by Hornbæk and Law (2007), who argue that usability studies should report such correlations in order to facilitate interpretation and comparison of usability evaluation outcomes. This section is therefore designed to highlight relationships between the most common usability measures: task success rate, time on task, participants' satisfaction with the targeted website (SUS), and the number of usability problems discovered, using Spearman's correlation test. Spearman's correlation coefficient is a statistical measure used to reveal associations between variables, to identify the strength of any correlations found, and to determine whether a correlation is positive or negative (Dewberry, 2004). Dewberry (2004) offers a guideline for interpreting the values of this correlation coefficient (*r*) to assess the strength of correlation:

(1) < 0.19: very weak

(2) 0.20-0.39: weak

(3) 0.40-0.59: moderate

(4) 0.60 -0.79: strong

(5) > 0.80: very strong

Table 4.37 sets out the results obtained from employing Spearman's correlation test, as follows:

- There is a statistically significant relationship between time spent on tasks and the number of usability problems discovered in the CTA and the HB conditions, suggesting that the participants who spent more time on tasks were able to discover significantly more usability problems. However, this was not the case for the RTA condition.
- There is no statistically significant relationship between problems discovered and participant satisfaction with the website, suggesting that finding usability problems did not affect satisfaction.
- There is, interestingly, no statistically significant relationship between task performance measures and participant satisfaction in any of the TA conditions. These findings are in line with previous research (Frøkjær et al., 2000; Hornbæk and Law, 2007; Nielsen and Levy, 1994b), which find low correlations between user performance and user satisfaction measures.

Table 4.37: Correlations amongst usability measures (N=20)

Usability measures		Task success	Task time	SUS	Usability problems
Task success	CTA	1	.223	.164	-.109
	RTA	1	-.157	.351	-.363
	HB	1	.114	.232	-.260
Task time	CTA		1	-.101	.494*
	RTA		1	-.217	.246
	HB		1	-.262	.512*
SUS score	CTA			1	-.212
	RTA			1	-.401
	HB			1	-.435
Usability problems	CTA				1
	RTA				1
	HB				1

* Correlation is significant at the .005 level (2-tailed).

4.10 Discussion

No previous study has investigated the effect of using two TA usability testing methods on the same user interface and compared this with another method. The present study has

found significant similarities and differences between CTA, RTA, and HB methods. This section presents the study's main findings, which are summarised in table 4.38, compares them to related research, and discusses the limitations of the study.

4.10.1 Think-Aloud Methods and Participants' Task Performance

Verbalising thoughts while working did not affect participants' task performance; that is, whether or not a participant was asked to think aloud during a usability session did not lead to a change in their task success rate or time spent on tasks. Reactivity was therefore not evident here. This implies that the task performance data collected when using concurrent thinking aloud can offer an accurate representation of real-world use. If usability practitioners wish to portray user performance in the "real context of use", they can thus choose between the CTA or HB methods on one hand and the RTA method on the other. These findings both correspond with and contradict earlier work by van den Haak et al. (2004), who found no differences in task performance between CTA and RTA methods but did find that thinking aloud led to significantly greater task accuracy. One possible explanation for this discrepancy is that van den Haak's et al. (2004) study did not take steps to control the participants' individual differences by matching them as closely as possible between conditions, as was done in the current study. Participants' demographic variables may therefore have affected van den Haak et al.'s results.

4.10.2 Think-Aloud Methods and Participants' Experience

With regards to participants' satisfaction with the tested website, thinking aloud while performing tasks seemed to have no effect on the perceived usability of the tested website, as assessed via comparison with participants in the silent RTA condition. This finding indicates that it is valid to collect data regarding participants' satisfaction when using concurrent thinking-aloud testing, which is in line with the findings of Olmsted-Hawala et al. (2010).

As in van den Haak et al. study, the CTA and RTA participants in the current study appeared to have similar testing experience. Most measures of the Experience with the TA Test questionnaire yielded neutral to positive judgements for the two evaluation methods, as they also did for the HB condition. This implies that stress and awkwardness, described

in section 1.4 as a potential negative influence on the functionality of the testing conditions, did not play major roles in participants' experiences. Therefore, it can be said that the ecological validity of the protocols (i.e. participants being comfortable with each protocol) is ensured. Nevertheless, the HB participants did find the task of verbalising their thoughts in the retrospective phase more time-consuming than in the concurrent phase and in the other two conditions. Overall, the results suggest that while in none of the three methods was ecological validity under serious threat, usability test participants might favour the CTA or RTA method over the HB method.

4.10.3 Think-Aloud Methods and Usability Problems Identified

The study's results indicate that the CTA and HB methods outperformed the RTA method in terms of the quantity and quality of usability problems detected at both the individual and final problem levels. Although Ericsson and Simon (1993) suggest that both concurrent and retrospective data can benefit the richness of data collected, results from the present study do not support their claim. The benefits of the HB method were not as anticipated, considering the efforts required from the participants and the evaluator. It only enabled the detection of a few more final problems, and did so at the cost of participants' experience and the evaluator's time and effort.

At the individual problem level, participants in the CTA and HB methods detected a higher number of problems than those in the RTA method, which corresponds with Peute et al.'s (2010) study comparing CTA and RTA methods. It was also evident from the present study that the CTA and HB methods identified more minor problems and layout problems and elicited more problems from the verbalisation source than the RTA method. There were no significant differences found between the CTA and HB conditions in terms of the number, sources, severity levels and types of individual problems detected. The latter result conflicts with that of Følstad and Hornbaek's (2010) study, which indicated that the retrospective session in the HB condition encouraged participants to identify more problems. This may be because in the aforementioned study, the researchers used interventions to specifically elicit solutions from participants, while in this study no interventions were used. At the final problem level, the CTA and HB methods detected more verbalised minor problems relating to layout problems than the RTA method. While

the HB method did detect five more problems than the CTA method, these were all verbalised problems with low severity levels.

4.10.4 Think-Aloud Methods and Cost

No previous studies have compared the cost, whether temporal or financial, of employing different TA methods. The findings of this study reveal that the CTA method cost substantially less than the RTA and HB methods in terms of the total time and potential financial cost required by the evaluator to conduct testing sessions and identify usability problems. In accordance with Følstad and Hornbæk's (2010) studies, the present study demonstrated that combined data collection in the HB condition requires a substantial investment of time and money. The RTA method is slightly cheaper than the HB method, but is still considerably more expensive than the CTA method. As most studies tend to compare the cost of CTA and RTA methods to other type of evaluation methods such as the heuristic evaluation method (Martin et al., 2014; Hasan, 2010; Andreasen et al., 2007), no comparison with previous studies can be made.

4.10.5 Think-Aloud Methods and Sample Size Needed

With regard to the relationship between the sample size and the number of problems detected, the results of this study highlight two important issues. The first is that Nielsen's (2000) optimistic view that five participants will suffice to detect most usability problems is challenged by the usability tests conducted in this study. The magic number of five participants failed to achieve its purported outcome of identifying 85% of problems; in fact, the best performing five participants in the three methods could not detect more than 43% of the total number of problems, and the first five users in the three methods could not find more than 24% of the total number of problems. These results are in agreement with researchers who raise doubts about the validity of small sample sizes for usability testing. (Molich et al., 2004; Lindgaard and Chattratichart, 2007). It appears that the complexity of websites such as online libraries is much greater than the complexity of the systems used to derive Nielsen's (2000) model, and that it is helpful to use (considerably) larger samples than those suggested by Nielsen (2000). Specifically, discovering 85% of problems requires 34 CTA participants, 46 RTA participants, or 30 HB participants.

The second issue is that the RTA method required considerably more test participants than the CTA and HB methods, which produced similar outcomes to one another, in order to find an equal percentage of problems.

Table 4.38: Overview of the main findings of the classic think-aloud study

Results in terms of	The classic TA study
Task performance	
- Successful task completion	No difference between the three TA methods
- Task duration	No difference between the three TA methods
Participant experiences	
- The tested website	No difference between the three TA methods
- The TA method	HB was considered more-time consuming than the other methods
Usability problems	
- Individual problems	
Detection means	RTA proved less fruitful than CTA and HB
Source of problems	CTA and HB produced higher number of verbalized problems
Severity of problems	CTA and HB produced higher number of minor problems
Types of problems	CTA and HB produced higher number of layout problems
- Final problems	
Detection means	RTA proved less fruitful than CTA and HB
Source of problems	CTA and HB produced higher number of verbalized problems
Severity of problems	CTA and HB produced higher number of minor problems
Types of problems	CTA and HB produced higher number of layout problems
Unique problems	CTA: 13, RTA: 8, HB: 17
Methods Cost	
- Temporal cost	CTA required much less time than the RTA and HB methods
- Financial cost	CTA would require much less financial cost than the RTA and HB methods
Sample size needed	RTA required considerably more test participants than the CTA and HB methods to find 85% of the problems

4.10.6 Limitations and the Next Experiment

Although the methods used for this research provided a large amount of data, the measurements may not have been fully accurate. For example, although the time on task was intended to be objective, it was actually a subjective measure because it was the evaluator's responsibility to start and stop the timer, which may not always have been 100% accurate. It was decided that for the next experiment, the researcher would use Morae software (Version, 2015)[21] in order to record time spent on tasks more objectively and to

[21] https://www.techsmith.com/morae.html

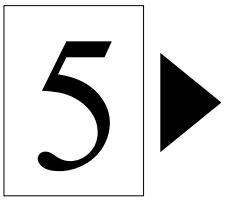
capture additional navigational behavioural data. This is discussed in section 5.7 in the next chapter.

4.11 Summary

This chapter has discussed the results of using the traditional think-aloud methods: the concurrent think-aloud method, the retrospective think-aloud method, and the hybrid method. These three methods were compared through an evaluation of a library website, which involved five points of comparison: overall task performance, test participants' experiences, quantity and quality of usability problems discovered, the cost of employing methods, and the relationship between sample size and the number of problems detected.

Overall, the findings revealed that the concurrent method can be argued to have outperformed the retrospective method and hybrid method in facilitating usability testing. It detected higher numbers of usability problems than the retrospective method, and produced output comparable to that of the hybrid method. The method received average to positive ratings from its users, and the possible reactivity associated with it was not observed in this study, as no differences between participants' task success rates were found for this method compared to the silent condition in the retrospective test. In addition, this method required much less time on the evaluator's part than the other two methods, which required double the testing and analysis time. Lastly, the concurrent and hybrid methods showed similar patterns in the relationship between sample size and the number of problems discover, and both outperformed the RTA method in this regard. These findings imply a basis for preferring the concurrent method over the retrospective and hybrid methods.

The next study will compare the performance of the classic concurrent think-aloud method with two relaxed variations of the method, wherein the evaluator played a more active role than in the traditional method.



RELAXED THINK-ALOUD STUDY

5.1 Overview

The previous chapter investigated the impact of Ericsson and Simon's (1993) classic think-aloud (TA) methods, namely the concurrent think-aloud method, the retrospective think-aloud method, and the hybrid method in usability testing. The results suggested that the concurrent think-aloud method was the most cost-effective method in collecting usability data. This chapter presents the second empirical study which explores the usefulness of two relaxed variations of the traditional concurrent think-aloud method, namely the active intervention protocol and the speech communication protocol. The chapter starts by stating the motivations behind the study, defines its specific aims, identifies the tested online library, and outlines the test tasks and participants. Following this, the chapter discusses the material and equipment used in the experiment, explains the experimental procedure, and presents the results of the pilot and main experiments. Finally, the chapter concludes by discussing and summarizing the results of the study.

5.2 Motivations

Despite the proven value of the traditional Concurrent Think-Aloud (CTA) method in assisting usability work, evidenced in the previous study and other related research (e.g., Peute et al., 2010), findings from field studies suggest that usability professionals often tend to adopt a more interactive approach - hereafter called the Active Intervention (AI) protocol - where practitioners intervene actively with the test participants during the TA process with questions asking participants for explanations and comments in the hope that it helps them to maximise the utility of the data produced (McDonald et al., 2012; Boren and Ramey, 2000). However, Ericsson and Simon (1993) emphasise the importance of minimal interaction between experimenter and participants, in order to guard against reactivity and evaluator induced bias. The difference between traditional CTA and the practice of usability professionals has caused some researchers to question whether another approach to thinking aloud might be more effective in usability studies than the classic method. Boren and Ramey (2000) proposed a theoretical alternative to the traditional TA protocol - referred to as Speech Communication (SC) protocol - where the evaluator plays the role of an active listener through the use of acknowledgment phrases to indicate to the participant that the evaluator is paying attention and is absorbed in the communication act, but no questions are asked and no conversation is made. Boren and Ramey (2000) considered their model a compromise approach between the AI approach, which risks

skewing the validity of collected data, and the traditional CTA technique which requests the evaluator to take the stance of a passive listener, which some usability professionals (and participants) may find inadequate, uncomfortable, or unrealistic (for more details on Boren and Ramey's model see section 2.6.3). To date, empirical studies have focused mostly on investigating the effect of using relaxed TA methods on participants' task performance and testing experience (Olmsted-Hawala et al., 2010; Zhao and McDonald, 2010; Hertzum et al., 2009). However, existing studies have yet to examine the impact of relaxed TA methods on the quantity and quality of usability problems discovered; the primary function of usability testing (Hartson et al., 2001) and nor has any study taken a holistic assessment of the methods.

5.3 Study Aims

The aim of this research exercise was to examine the utility and validity of two relaxed variations of the classic CTA protocol: the AI protocol often used in usability practice (McDonald et al., 2012), and the SC protocol put forward by Boren and Ramey (2000). This was achieved by comparing the two methods with the CTA method. The three methods were compared through an evaluation of a library website, which involved five points of comparison: overall task performance, test participants' experiences, quantity and quality of problems discovered, cost of employing the methods, and the relationship between the sample size and the number of problems detected in each condition.

5.4 Test Object

This research focuses on university library websites as test objects, due to the reasons given in section 3.7. It was not possible to use the website evaluated in Study One, as the website administrator could not confirm that the website interface would be stable during the timeframe of the current study. The process of selecting the targeted online library website for this study was based on the same criteria reported in Chapter 4 (section 4.3). This would maintain the validity of the research, and enable the results of the current study to be compared with the results of the previous one.

Out of several options, the Durham University (DU) library website[22] was deemed a suitable candidate for the experiment in this study (see Figure 5.1). Once the website was selected, the researcher contacted the website administrator via email (see Appendix D2) to obtain consent to use the site, and to establish in advance that there was no intention to modify or alter the interface either prior to, or during, the study. An attempt was also made to ensure that the selected website would be stable for a long period of time which would enable its use in the third study of the research (co-participation study). The administrator of the DU library website gave the researcher written consent (see Appendix D3) to evaluate their website and assurances that the interface would not be modified prior to or for the duration of the intended period for the current study or the expected period of the third study.

For clarity and simplicity throughout this chapter, the title (DU-L) is used to refer to the DU library website. As Figure 5.1 below shows, the DU-L website home page has a comprehensive search tool positioned in the middle and a number of links for various options that are standard to most academic libraries' websites: conducting searches, borrowing and reserving items, finding subject information, etc. All information on the site was only available in English.

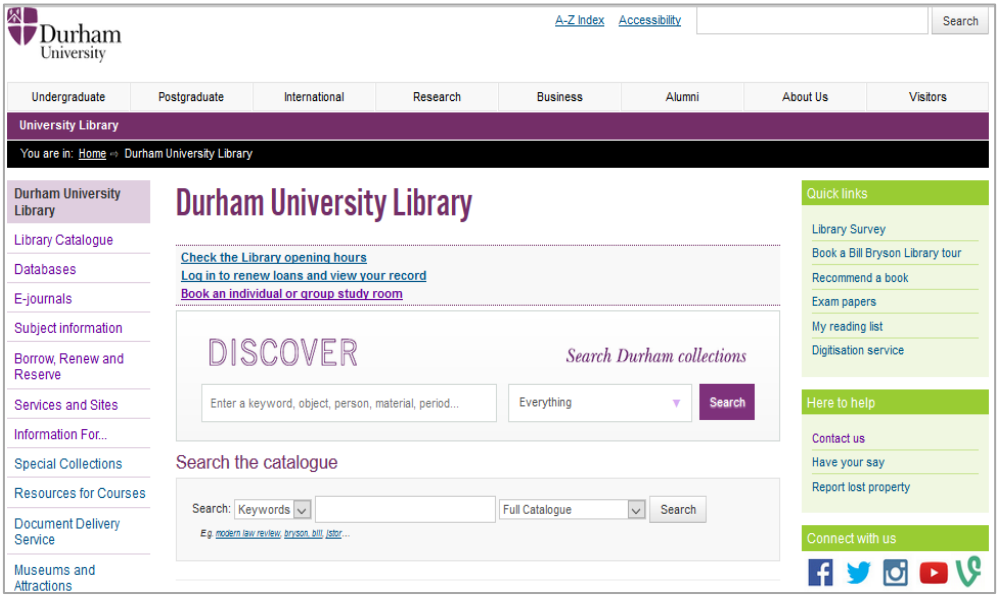


Figure 5.1: Screenshot of the test object's homepage

[22] https://www.dur.ac.uk/library/

After the test object was defined, a series of tasks was designed to evaluate the selected website by these three differing TA methods.

5.5 Tasks

As in the previous study, a context of use analysis of the chosen website was conducted (Maguire, 2001) to identify its intended audience and the most common tasks they undertake on the site. Furthermore, the selected website was evaluated by the author and an additional usability expert using Heuristic Evaluation (Nielsen, 1993a) to identify potential problematic areas, which would provide the focus for the task design.

At the website administrator's request, the context of use analysis question list (see Table 4.1 in) was sent to him via email. Table 5.1 in the next section summarises the information gained from this analysis. The administrator stated that the users of the website mainly used the site to search the library catalogue, use e-resources list, use databases list, book study rooms, find borrowing information and look up opening hours. The site's interface is mainly accessible on desktops/laptop browsers. It is also reachable via mobile browsers, though with limited usage. The administrator mentioned that Durham University had carried out very basic evaluations on the website, and the results revealed that the site suffered from a number of usability issues, including but not limited to, navigational problems, overuse of jargon, inconsistency issues, and content and layout problems.

The researcher used the information acquired from the website's administrator and the results of the heuristic evaluation to create nine different scenario tasks that together covered the targeted website's main features and predicted problematic areas: finding borrowing information (Task 1), finding information regarding off-campus services (Task 2), booking a study room (Task 3), searching the library catalogue using its simple search (Task 4), searching the library catalogue using the advanced search (Tasks 5-9) (see Appendix D4). All tasks were designed to be carried out independently from one another, meaning that even if a task was not completed successfully, participants could still carry out the other tasks. The tasks were piloted with three people prior to the commencement of data collection. An example task is shown below:

> 'Task #2: You are a part-time student who works off-campus most of the time. You want to know what services the library offers for off-site users. Can you find them?'

Once the experimental tasks had been designed, the recruitment phase began. This process will be discussed in the following section.

5.6 Participants

The number and background of test participants are key factors in the selection process (Sauro, 2010). As in Study One, it was decided that 60 participants would be recruited for the main experiment in this study, with 20 participants being allocated to each TA condition. It was also determined to recruit three additional subjects to conduct the pilot experiment, and another three individuals to cover for no-shows. This made the desired sample size for all components of the study 66 participants.

According to the site administrator, the library site is mainly intended for students and academic staff at DU, although it can also be accessed by staff at other establishments, and visitors, who together represent its secondary users. The administrator reported a lack of information regarding detailed demographic details of the sites' users, but he stated that "the assumption is that they would roughly follow the data on the University population"[23] which indicates that the student and teaching staff come from different cultural backgrounds, with British being the majority, and from a wide range of academic levels and areas of study. Since the primary and dominant users of the tested website are students, it was decided that all participants in this study would be university students (further justification is provided in section 4.5). This means the recruiting criteria for this study are in line with the ones applied in the previous study (see Table 4.3 in Chapter Four).

Table 5.1: Results of the context of use analysis

Users	Main task goals
Primary users	To search the catalogue
Students	To use e-resources list
Academic Staff	To use e-databases list
Secondary users	To book study rooms
Other staff	To find out borrowing information
Visitors	To look up library opening times

Following the recruitment process outlined in Chapter Four (section 4.5), sixty UEA students who were applicable with the recruiting criteria were invited via email to

[23] https://www.dur.ac.uk/student.registry/statistics/summary/1.1summary/

participate in the main study. Three students who were almost in line with the recruiting criteria were assigned for the pilot study, and another three students were invited as back-ups to offset no-shows. Section 5.10.1 provides more details regarding the participants in the main study.

5.7 Setting and Equipment

All experimental sessions were conducted in the same laboratory in the School of Computing sciences at UEA. Two computer laptops were used in the experiments. One laptop was used by the participants to navigate the website, and the other was used by the evaluator to observe the participants' screen. The two laptops were connected with a wire network (see Figure 5.3). The computer laptop and Internet browser the participants used was the same used in Study One. The Morae (Version, 2015)[24] software package was used in all the experiments to record the whole test process. The researcher decided to use Morae software in this experiment in order to record time on task more objectively, and to capture additional navigational behaviour data, as discussed below.

Morae is a software-based solution for usability testing, which enhances data collection and speeds up analysis. It consists of three software parts (Morae recorder, Morae observer and Morae manager). The three parts work together to provide a complete picture of the testing. With the Morae recorder, the screen and the navigational behaviour data of the participant such as mouse clicks and pages visited which can offer better insights into how TA methods affect task performance, the faces of the participant and the evaluator (through a web camera) and the audio of the participant and evaluator (through a microphone) can be recorded at the same time. It was installed on the participants' computer laptop. The recorder runs silently in the background, and when it starts to work (pressing the red button) it will become a small icon on the right corner; however, most people will not notice this and it does not disturb users. With Morae Observer (Figure 5.2), installed on the evaluator's computer, the evaluator can observe the interaction of the user with the screen, record the observations, take notes and record other relevant matters. Figure 5.2 shows what the researcher can see through the Morae observer. Morae Manager was used later in the analysis to review the session videos.

[24] https://www.techsmith.com/morae.html

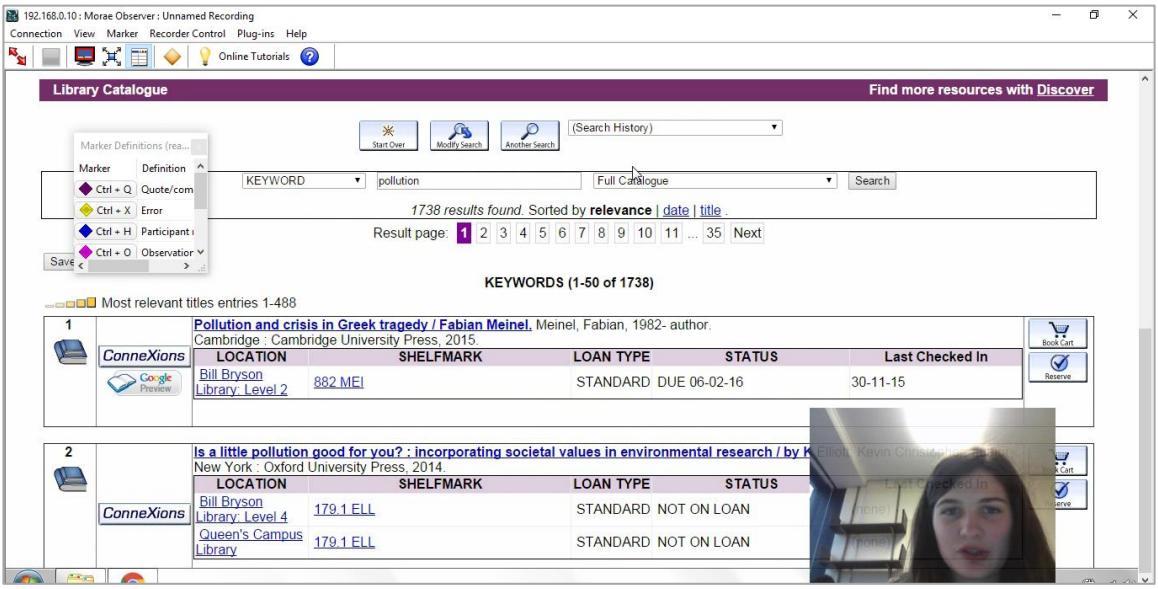


Figure 5.2: Morae observer (picture taken with participant's permission)

Figure 5.3: Equipment used

5.8 Experimental Procedure

Permission to run the study was sought and granted from the University's Ethics committee (see Appendix D1). Each testing session was conducted on a one-to-one basis, with only the evaluator and participant present at a time. Except for the level of their interaction with the evaluator, all testing sessions followed the same procedure. A graphical representation of the procedure is given in Figure 5.4. The session began with the researcher welcoming each participant and asking them to read and sign a consent form (see Appendix D5).

Participants were informed that the purpose of the study was to evaluate an online library website.

Each participant was then given a maximum of two minutes to familiarise his or her-self with the lab computer. Following this, the concept of thinking aloud was introduced using Ericsson and Simon's instructions (1993). Regardless of their TA condition, the same basic instruction on the TA technique was used. This guideline was taken from Ericsson and Simon (1993, p.376). Participants were instructed to think aloud while performing the tasks and not to turn to the evaluator for assistance; they were however informed that if they did fall silent for a period the evaluator would ask them to keep thinking aloud. Participants received both verbal and written instructions to do so (see Appendix D2). The participant then engaged in a brief think-aloud practice session using the simple and neutral task of looking up the word "*chant"* in an online dictionary.

On completion of the training session, the participants were asked to read the task instructions shown on the screen, before beginning task solving. The participants were instructed to choose the "Start task" option in Morae recorder when they were ready to begin the task and the "End task" option once they believed they had retrieved the required information, or if they recognized they were unable to find any appropriate information. Morae recorder displayed the test tasks in a counterbalanced order to prevent the order in which the tasks were presented from affecting the results (Sauro, 2010).

During participants' task performance, the evaluator remained in the same room as the participants and, was seated a short distance behind the participant on their right hand side. For the *traditional CTA condition*, Ericsson and Simon's guidelines were strictly followed; the only interaction between the evaluator and the participants was to issue the "please keep talking" reminder if participants had fallen silent for 15 seconds. For the *SC condition*, the evaluator followed the TA technique proposed by Boren and Ramey (2000); using acknowledgement tokens in form of the affirmatory "Mm hmm" with intonation, and probing with tokens of "Mm hmm?" or asking "And now…?" if participants fell silent for more than 15 seconds, and if the former questioning tone failed to elicit response. For the *AI condition*, the evaluator intervened actively with participants. Zhao and McDonald (2010) developed a list of interventions in their comparative study of the AI method and the traditional CTA method which was mostly based on the recommendations of the

authors of usability textbooks that suggested intervening during usability testing (e.g., Nielsen, 1993a; Dumas and Redish, 1999). This researcher and his supervisor also set out a Master's project to explore the types of interventions usability practitioners used in the practice. The project used a combination of questionnaires filled in by 47 usability practitioners and an observation of a professional usability company in London in 2015 (Naveedh, 2015). The project results were utilised alongside the information acquired from the relevant literature (Zhao and McDonald, 2010) to determine how the researcher would intervene with participants in the test sessions (e.g., the evaluator asks direct questions about different areas of the website where the participant is having difficulty or is describing an area as confusing or frustrating). For the full list of interventions types and associated triggers see Appendix D7.

When participants had completed the tasks, they were asked to fill in the two online post-test questionnaires to provide feedback on the evaluated website (the System Usability Scale - SUS - questionnaire) and the test (experience with TA test questionnaire). Lastly, the evaluator thanked each participant for taking part and gave them £5 as token of appreciation for participating in the study.

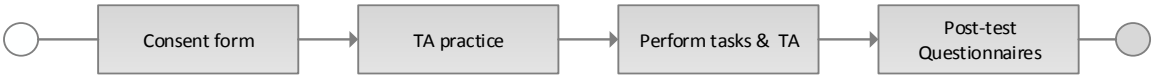


Figure 5.4: Experimental procedure

5.9 Piloting and Correction

The literature (see section 2.4) and previous study showed that pilot tests are an important and valuable tool for the experimenter. Three sessions of pilot studies were therefore conducted with UEA students before the actual experiment in this study. In addition to identifying potential methodological issues, piloting also served to ensure that the author was familiar with the interventions that would be used in the AI condition. The pilot study was conducted well in advance of the main study to allow time for any necessary action to be taken to address issues which might arise. In short, pilot participants were asked to think aloud whilst using the targeted website to perform the tasks they had been allocated and were given the opportunity to ask questions before commencing each task so that any unclear task wording could be identified. Some minor changes were made to the wording of tasks as a result, in order to ensure maximum clarity. Other aspects of the pilot tests

went smoothly and remained part of the formal test procedure. The subsequent sections explore the findings obtained from the main study.

5.10 Results

This section presents the results obtained from employing the three TA testing methods (CTA, SC, and AI) used in this study. It starts by outlining the characteristics of the participants assigned to the three TA conditions (subsection 5.10.1). It then presents the results for participants' task performance (subsection 5.10.2), participants' experiences (subsection 5.10.3), usability problems discovered (subsection 5.10.4), the cost of employing the methods (subsection 5.10.5), the relationship between sample size and usability problems discovered (subsection 5.10.6), before concluding with a correlational analysis of the usability measures used (subsection 5.10.7).

5.10.1 Participants' Profiles

Table 5.2 summarises the demographic profile and descriptive statistics of the participants. As shown in the table, 60 UEA students participated in this study; 39 (65%) of whom were male and 21 (35%) were female. 65% of the participants were male and 35% were female as well in each condition, a 13/7 split. It was challenging to recruit female participants in this study. A possible reason could be the skewed male/female ratio in the representative sample composition. In any case, skewed ratios are a common problem in voluntary surveys. This problem is known as the "self-selection bias" in which some participants are more likely to participate in the survey than others (Rubin and Babbie, 2009). Due to the higher self-selection tendencies of male participants, and the resulting shortage of female respondents, the final sample had a male/female ratio of 1.66 (35 male and 21 female participants). Although this is a skewed sample, it was not considered to have an adverse impact on the comparative results as the TA groups had the same number of male and female participants.

While a number of participants were from European countries (9) and North America (4), the majority (47) were British. The few students for whom English was not their first language rated themselves to be excellent at reading and speaking English. Additionally, the researcher ensured that they had passed IELTS (International English Language

Testing System, above 6.5 points) or any other established English proficiency tests with a score accepted by UEA (Shi, 2009) in order to mitigate the impact of language proficiency on TA behaviour (Sun and Shi, 2007). The participants selected were all in the age category of 18-39, 71.66% were 18-29 years old, and 28.33% were 30-39. Only 5 out of 60 participants had previously taken part in a usability study, and not recently (i.e. last six months). An attempt was made to assign these individuals evenly to the groups.

All participants were frequent users of the Internet, and had all visited online library sites before, but none had visited the site used in this study. By being part of the target group (i.e. university students) as well as novice users of the targeted website, the participants were very suitable for evaluating the DU-L website. A Kruskal-Wallis H test was run in order to statistically determine if there were significance differences in participants' demographics between the TA groups. The distributions of nationality ($\chi^2(2)= 0.804$, $p= .669$), gender ($\chi^2(2)= .000$, $p= 1.00$), age ($\chi^2(2)= 3.27$, $p= .194$), and Internet use ($\chi^2(2)= 4.37$, $p= .112$) were similar for all groups. Accordingly, it may be said that the participants' demographics did not impact the results.

Table 5.2: Summary statistics of demographic characteristics of participants

Characteristics		**CTA** (n=20)	**SC** (n=20)	**AI** (n=20)	**Total** (n=60)	**Percent**
Country	Britain	15	15	17	47	78.33
	European	5	2	2	9	15
	America	0	3	1	4	6.66
Gender	Male	13	13	13	39	65
	Female	7	7	7	21	35
Age	18-29	11	16	15	42	70.00
	30-39	9	4	5	18	30.00
Internet use	Daily	18	16	20	54	90
	At least once a week	2	4	0	6	10

5.10.2 Task Performance

Task performance measures are used to assess the possible reactivity associated with TA methods: a change in task performance due to the double workload of having to perform tasks and think aloud simultaneously (Fox et al., 2011). The task performance of participants in the three TA conditions were measured in this study using four indicators:

task completion rate, time on task, mouse clicks, and number of pages browsed. Since it was evident from Study One that classic CTA has no impact on task performance, the CTA group were regarded as the control group in this study; that is, results from the other two groups (SC and AI) would be compared against the results from CTA participants. The following sub-sections show the results of the performance measures.

5.10.2.1 Task Completion

Each participant was asked to perform nine tasks on the targeted website, meaning that a total of 180 tasks were performed by each group. Participants in the CTA group successfully completed 110 tasks out of 180 tasks (61% success rate), the SC participants were able to complete 106 tasks (58% success rate), and the AI participants completed 101 tasks (56% success rate). In other words, participants in the CTA group completed an average of 5.50 out of the nine tasks, in contrast to an average of 5.30 tasks completed by participants in the SC group, and 5.05 tasks completed by the AI group (see Table 5.3). Therefore, participants in the CTA condition had the highest completion rate and participants in the AI had the lowest completion rate. The inferential statistics presented in Table 5.4 will provide a better indication of the differences in the means and the significance of those differences. The most difficult task (Task 6) was completed successfully by only twenty seven of the sixty participants. In contrast, the easiest task (Task 1) was completed by a vast majority: 55 out of the 60 participants.

Table 5.3: Descriptive statistics of the task completion for the TA methods

Task completion	CTA	SC	AI
Total number of tasks	180	180	180
# of successful tasks	110	106	101
Percent of successful tasks	61%	58%	56%

To determine the level of variance between the samples and to understand whether that difference in the total number of successful tasks is statistically significant, a one-way ANOVA test was run. As mentioned in section 4.9.2, the one-way ANOVA is a parametric test used to compare the means of three or more unrelated groups, and assumes the approximate normal distribution of the data, and the homogeneity of variances (Filed, 2005). For data distribution to qualify as approximate normal, the *p*-value of the Shapiro-Wilk test must be more than 0.05 for each group of the independent variable. To meet the

assumption of homogeneity of variance, the *p*-value of the Levene's test must be more than 0.05.

Task success rates were approximately normally distributed for the three TA groups as verified by Shapiro-Wilk test with *p*=.076 for the CTA group, *p*=.188 for the SC group, and *p*=.378 for the AI group, respectively. The second assumption of the ANOVA test was also met as there was homogeneity of variances (*p*=.253). A one-way ANOVA test with alpha=.05 found no significant difference in the number of successful task completions between the three TA conditions, as shown in Table 5.4.

Table 5.4: Inferential statistics of the task completion for the TA methods

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Task success	5.50	1.59	5.30	1.39	5.05	1.05	F(2,57)=.62, *p*= .537

5.10.2.2 Time on Task

As the name suggests, this measure quantifies the time that participants spent on the test tasks. For each TA condition, the time that participants spent on the test tasks, regardless of whether the tasks were completed successfully, was calculated. Table 5.5 compares the total time spent on all tasks by all participants and the mean time spent on tasks.

Table 5.5: Descriptive statistics of time on tasks for the TA methods

Time on task	CTA	SC	AI
Overall time spent on tasks (*m*)	503	555	624
Mean time spent on tasks (*m*)	25.15	27.75	31.20

Examining these results reveals that the participants in the AI condition took longer to complete the tasks compared to the participants in the CTA and SC conditions. The AI group spent a total of 624 minutes on tasks, whereas the CTA and SC group spent a total of 503 minutes and 555 minutes, respectively. In other words, participants in the AI group had an average of 31.20 minutes on the nine tasks, in contrast to an average of 25.15 minutes by the CTA group, and 27.75 minutes by the SC group (see Table 5.5). A one-way ANOVA test was conducted in order to determine if there were statistically significant differences in the mean time spent on all tasks. The Shapiro-Wilk test showed that the task time were approximately normally distributed for the three think-aloud groups, with *p*= .099 for the CTA group, *p*= 0.181 for the SC group, and *p*= 0.293 for the AI group,

respectively. The assumption of homogeneity of variances was also met (p= 0.561). The one-way ANOVA test found significance difference among the three verbalization conditions, as shown in Table 5.6. The researcher conducted a Tukey post-hoc analysis to determine which condition had the biggest effect on task time, and which condition was significantly different from the control condition (i.e., CTA condition). The post hoc analysis revealed that participants in the AI condition worked significantly slower on tasks than the CTA ($p<0.0001$) and SC ($p< 0.05$) participants did (see Table 5.6). The prolonged task completion in the AI condition might be attributed to four reasons: first, it was merely due to the additional dialogue between the participants and the evaluator which slowed down the process. Second, the evaluator's interventions might disrupt the participants' mental processes and made them less able to focus. Third, it made them doubtful about their approach to solving tasks and pushed them to redo some interactions with the system. Fourth, the active interaction between the evaluator and the participants in the AI condition might situate the participants in a more social environment. This might consequently encourage them to try harder in performing tasks and explore more solution paths in order to impress the evaluator. However, the absence of differences in the number of correctly solved tasks does not seem to lend support to the last explanation. Investigating the navigational behaviour measures and participant test experience would further reinforce or repudiate these explanations.

Table 5.6: Inferential statistics of time on tasks for the TA methods

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Time on tasks (*min*)*	25.15	3.45	27.75	3.78	31.20	4.88	$F(2,57)=11.03$, $p<0.0001$

* AI differed significantly from CTA ($p< 0.0001$) and SC ($p<0.05$)

5.10.2.3 Navigational Behaviour

As mentioned in section 5.8, Morae software records a variety of navigational actions such as mouse clicks and browsed pages. Such data can offer greater insights into the influence of TA methods on user behaviour (Hertzum et al. 2009). It can also assist in understanding the efficiency of a particular website or application (Tullis and Albert, 2008). To examine if there is a significant difference in the navigational behaviour measures between the TA conditions, one-way ANOVA test needed to run. Table 5.7 shows the results of the normality test and the homogeneity of variance test for the navigational behaviour data. It can be clearly seen that the values of the normality and homogeneity of variance tests for

the mouse click variable are larger than 0.5, so it can be claimed that the assumptions of one-way ANOVA were met. However, for browsed pages the assumption of homogeneity of variances has been violated. As mentioned in section 4.9.2, if the sample size in each group is similar, violation of this assumption tends not to be a serious issue (Filed, 2009). As part of one-way ANOVA procedure SPSS produces a table that includes the *p*-value when the assumption of homogeneity of variances is met and another *p*-value when the assumption is not fulfilled. The statistical result reported here was based on equal variances not assumed.

Table 5.7: Tests for normality and homogeneity of variance for the navigational measures

	Shapiro-Wilk test			**Levene's test**
	CTA	**SC**	**AI**	
Mouse clicks	.638	.501	.722	.515
Browsed pages	.371	.279	.163	.040

A one-way ANOVA test showed that there was a statistically significant difference between the conditions in the overall number of mouse clicks and pages browsed. The Tukey HSD post-hoc analysis revealed that the AI participants clicked their mouse significantly more, and visited more pages, than the CTA and SC participants (see Table 5.8). The increase in navigational behaviour during AI condition further lend support to the idea that evaluator's active interventions may disrupt the participants' mental activities and make it more difficult to maintain a focus, and possibly necessitate they redo some interactions with the system. Another reason could be that AI made participants doubtful about their approach to solving tasks, or cognisant of other ways of solving them, leading to more navigational exploration of the website.

Table 5.8: Navigational measures for the TA methods

	CTA		**SC**		**AI**		**Value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Mouse clicks*	105.20	22.70	109.25	29.25	125.00	25.00	$F(2,57)=4.14$, $p= .021$
Browsed pages*	34.80	7.86	37.30	8.74	43.55	14.60	$F(2,57)=6.22$, $p= .004$

* $p< 0.05$ significance obtained

In all, the results in this section suggest that probing participants with questions while performing tasks and thinking-aloud prolongs task performance and affects navigational behaviour. The next section will present the rating of participants regarding their test experience.

5.10.3 Participants' Experiences

Apart from the task statistics, the participants' experiences with the process of usability testing can also serve as an important indicator of the success or failure of the process. To gauge this, participants were asked to fill out two post-test questionnaires. The first one (SUS questionnaire) dealt with their satisfaction with the usability of the chosen website (Appendix B2), while the second one (experience with TA test questionnaire) dealt with their experiences with the testing process (Appendix B1).

5.10.3.1 Participants' Satisfaction with the Usability of the Website

The SUS form consists of 10 questions to be answered on a 5-point Likert scale with 1 indicating strong disagreement and 5 indicating strong agreement. SUS yields a single score on a scale of 0–100 representing the overall usability of the website (Brooke, 1996). The higher the score, the more satisfied the participant reported being with the site. The analysis reveals that the three participant groups did not find the system usable. The scores are all below the average SUS score of 68 established by Nathan Thomas (2015). The CTA condition gave the highest score, while the AI condition gave the lowest score. Having met the assumptions of normality (p= 0.448 for the CTA group; p= 0.137 for the SC; and p= 0.653 for the AI group) and homogeneity of variances (p= 0.745), a one-way ANOVA test was conducted, and indicated that the satisfaction rating did not differ significantly between the conditions (see Table 5.9).

Table 5.9: Participants' satisfaction with the usability of the tested website

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
SUS score	61.60	10.58	58.55	13.37	56.40	15.82	F(2,57)=.40, p=.670

On a totaled scale of 1 to 100

5.10.3.2 Participant Experience with the TA Test

The post-test questionnaire, related to participant experiences with the test, consisted of ten quantitative questions to be answered on a 5-point Likert scale with 1 indicating strong disagreement and 5 indicating strong agreement. The questions concentrated on the testing process itself to gauge the ease of performing it. The participants were asked to rate their experience with: (1) how the TA method affected normal working on tasks; (2) having to think aloud concurrently; and (3) the presence of the evaluator. Since the normality tests show that there is a departure from the norm for many variables (see Appendix D10), a

non-parametric test, the Kruskal Wallis H test, was used to analyse the data. Table 5.10 presents the results of participants' ratings in the three TA conditions.

To begin, all participants were asked to estimate how their working procedure on the experimental tasks differed from their normal working, by marking on a five-point scale their perceived speed and focus differential whilst involved in the study. A Kruskal Wallis H test and Bonferroni post-hoc analyses indicated that participants in the AI condition felt they worked significantly slower when thinking aloud than participants in the CTA ($p< 0.005$) and SC ($p< 0.05$) conditions. These results are in line with the data about performance and navigational behaviour, and support the claim that the participants' task performances were clearly affected by the evaluator's active probing. The participants in the CTA and SC conditions indicated that they had not worked all that differently from usual with average scores ranging from 2.50 to 2.95.

Participants were also asked to indicate to which degree they thought having to think aloud was difficult, unpleasant, tiring, unnatural, and time consuming. The results showed that there were no significant differences between the methods. On average, the participants rated their experiences with thinking aloud neutrally, with scores ranging around the middle of the five-point scale.

The third and final part of the questionnaire involved questions about the presence of the evaluator. Participants were asked to indicate to what degree they found it unpleasant, unnatural or disturbing to have the evaluator present during the experiment. A Kruskal Wallis H test and Bonferroni post hoc analyses revealed a significant difference in the level of distractions caused by the evaluator: Participants in the AI condition felt more distracted than their colleges in the other two conditions. No differences were found in other aspects. This difference can again be explained by the active intervention of the evaluator. The AI participants had to actively perform tasks and TA, and at the same time answer the evaluator's questions which made the test situation considerably more distracting than in the CTA and SC conditions. With all scores ranging from 1.10 to 1.60, the CTA and SC participants clearly felt that they were not affected by the presence of the evaluator.

Table 5.10: Participants' experience with the TA test

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Working condition							
Slower than my normal working*	2.50	1.19	2.70	1.41	3.85	1.34	$\chi2(2)=10.1$, $p=.006$
More focused than my normal working	2.70	1.36	2.95	1.79	3.05	1.31	$\chi2(2)=1.09$, $p=.579$
Think-aloud experience							
Difficult	2.10	1.07	2.30	0.73	2.55	1.31	$\chi2(2)=2.81$, $p=.245$
Unnatural	2.85	0.44	3.00	0.50	3.25	0.51	$\chi2(2)=1.81$, $p=.403$
Unpleasant	2.45	1.14	2.30	1.59	2.70	1.38	$\chi2(2)=1.32$, $p=.516$
Tiring	2.20	1.00	2.00	1.12	2.60	1.63	$\chi2(2)=1.29$, $p=.524$
Time-consuming	2.60	1.45	2.60	1.42	3.00	1.54	$\chi2(2)=1.89$, $p=.387$
Evaluator presence							
Unnatural	1.50	0.93	1.35	0.67	1.65	0.90	$\chi2(2)=1.30$, $p=.520$
Disturbing**	1.45	1.17	1.60	0.88	2.70	1.71	$\chi2(2)=17.0$, $p<0.0001$
Unpleasant	1.25	1.23	1.10	0.44	1.40	1.23	$\chi2(2)=3.03$, $p=.219$

Five-points scale (1: Strongly disagree to 5: Strongly agree) * p< 0.05 significance obtained ** p< 0.005 significance obtained

In summary, the results indicated that the AI method is evaluated least positively by its users. The other two test approaches revealed similar results with regard to the participants' experiences. This finding suggests that usability test participants prefer using the CTA method or the SC method over the AI method.

5.10.4 Usability Problems

This section focuses on the quantity and quality of the problems detected per participant (i.e., individual problems) and in each TA condition (i.e., final problems). A non-parametric Kruskal Wallis H test was used for the analysis of the individual problem data because the data were not normally distributed (see Appendix D11), which is normally the case in usability tests (Dumas and Redish 1999). Descriptive statistics were used to describe and summarize the final problems discovered.

5.10.4.1 Individual Usability Problems

Table 5.11 gives an overview of the mean number of problems detected per participant in each TA condition. In the table, a distinction is also made according to the way the problems had surfaced: (1) by observation; (2) by verbalization; or (3) by a combination of observation and verbalization (for problem source coding details see section 4.9.4.1). Interestingly, Kruskal Wallis H testing revealed that there were no significant differences

between the three TA testing variations, either in terms of the number of individual problems detected or in terms of the ways in which these were detected. Therefore, the additional interaction between the evaluator and the participants in the relaxed conditions did not seem to maximise the utility of the data produced. The most interesting outcome is that the results of AI condition showed no significant differences, compared to the CTA and SC conditions. As such, the fact that the evaluator in the AI was intervening with the participants during the TA process did not cause the participants to detect a significantly larger number of problems than participants in the other two conditions. One possible explanation for this result could be that the AI participants might have considered some issues to be obvious, therefore not worthy of further explanation and reporting. Participants possibly felt their task performance was distracted by the evaluator and this might have caused them to give more priority to task performance and discouraged them from responding fully to the evaluator questions. Alternatively, the psychological effect of probing the participants with questions might make some participants feel they were not contributing as expected and may have put them in a "novice-expert" mode which made them feel reserved and uncertain about sharing additional information about the usability issues of the site.

Table 5.11: TA methods and the number of individual problems

	CTA		**SC**		**AI**		**Value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Observed	2.50	2.06	2.25	1.86	3.10	1.73	$\chi2(2)= 3.09$, $p= .213$
Verbalised	2.20	1.28	2.40	1.53	2.85	2.41	$\chi2(2)= .44$, $p= .978$
Both	6.60	3.78	6.30	2.93	7.05	2.83	$\chi2(2)= .117$, $p= .555$
Total	11.30	3.96	10.95	3.79	13.00	4.13	$\chi2(2)= 3.70$, $p= .157$

Individual usability problems and severity levels

Individual problems were also coded according to severity level to four types: 1) critical, 2) major, 3) minor, and 4) enhancement (for problem severity coding details see section 4.9.4.1). Table 5.12 presents the mean value and the standard deviation of the number of problems detected for each of the severity levels. A Kruskal Wallis H test with Bonferroni post-hoc analyses found a significant difference between the methods regarding the number of individual problems belonging to the severity level of enhancement. The AI method produced more enhancement individual problems than the CTA and SC methods, but this difference concerned only a very small number of problems (0.25 and 0.15 as

opposed to 0.7). There were no differences between the methods for the number of individual problems classified as critical, major, or minor.

Table 5.12: TA methods and individual problem severity levels

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Critical	3.50	0.94	3.55	0.75	3.85	0.70	$\chi 2(2)= 2.11$, $p= .348$
Major	4.20	1.50	4.35	2.00	4.80	1.85	$\chi 2(2)= .793$, $p= .673$
Minor	3.35	2.45	2.90	1.86	3.65	2.20	$\chi 2(2)= 1.63$, $p= .442$
Enhancement*	0.25	0.55	0.15	0.36	0.70	0.62	$\chi 2(2)= 11.0$, $p= .004$

* p< 0.005 significance obtained

Individual usability problem types

To investigate the types of problem that were detected in the three conditions, two independent usability experts divided all detected problems into four specific problem types: navigation, layout, content, and functionality (for problem type coding details see section 4.9.4.1). The inter-coder reliability was computed using Cohen's kappa (explained in section 4.9.4.1). The overall kappa was 0.79, which indicates a highly satisfactory level of inter-coder agreement.

Table 5.13 shows the overall distribution of problem types in the three methods. As in the previous experiment, all participants clearly experienced most difficulties in navigating the website and interacting with its layout. The results for the other problem types were quite similar across the three conditions too, with only one significant differences between CTA and SC. The CTA and SC conditions differed in respect to content. However, these differences were only slightly significant ($p<0.05$). As follows, the three conditions largely revealed similar types of problems in similar frequencies.

Table 5.13: TA methods and individual problem type

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Navigation	4.45	1.57	4.30	1.49	5.05	1.60	$\chi 2(2)=3.09$, $p=.213$
Layout	4.00	1.86	3.80	1.70	4.50	1.96	$\chi 2(2)=1.76$, $p=.414$
Content	0.65*	0.48	0.25*	0.55	0.40	0.50	$\chi 2(2)= 6.54$, $p=.038$
Functionality	2.20	1.07	2.60	1.23	3.05	1.79	$\chi 2(2)= 3.80$, $p=.149$

* p< 0.05 significance obtained

5.10.4.2 Final Usability Problems

In total, 98 problems were extracted from the test session files of the three conditions (Table 5.14). The CTA condition generated 60 problems (61%), 16 of which were unique to that condition, the SC condition yielded 58 problems (59%), 12 of which were unique to that condition, and the AI condition produced 64 problems (65%), 19 of which were unique to that condition. Overall, these results are in line with the result of the number of individual problems detected and thus reinforce the idea that each of the three methods is equally fruitful in terms of the quantity of detected problems.

Table 5.14: TA methods and the number of final problems

	# of problems	% of problems	# of unique problems	% of unique Problems
CTA	60	61 %	16	16 %
SC	58	59 %	12	12 %
AI	64	65 %	19	19 %
Total	98	100 %	47	47 %

There were 33 (33%) problems that occurred in each of the three conditions. The overlap between two rather than three conditions was substantially less, ranging from 5% to 8%. These low percentages indicate a substantial number of unique problems identified by each of the three conditions (47 problems). The Venn diagram in Figure 5.5 shows the overlap between the three TA protocols. Appendix D9 lists the final problems discovered by the participants in this study.

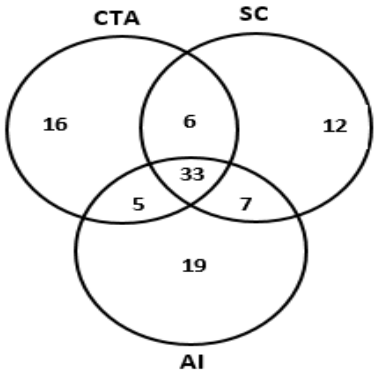


Figure 5.5: Venn diagram showing overlap in problems between think-aloud protocols

Final usability problems and their sources

The final usability problems were coded according to verbalisation source, observation source, and a combination of both, as explained in section 4.9.1.2. The results are shown in Table 5.15. As shown in the table, in the CTA condition, 7 problems were derived from observation evidence, 20 from verbal evidence and 33 from a combination of the two. For the SC condition, 5 problems were derived from observation evidence, 18 from verbal evidence and 35 from a combination of the two. For the AI condition, 8 problems were derived from observation evidence, 21 from verbal evidence and 35 from a combination of the two. In terms of the unique final problems, the vast majority of unique problems in the CTA (75%), the SC (83%), and the AI (79%) conditions came to light from the verbalization source.

Table 5.15: TA methods and final problem sources

	CTA		**SC**		**AI**	
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**
Observed	1	6	0	5	3	5
Verbalised	12	8	10	8	15	6
Both	3	30	2	33	1	34
Total	16	44	12	46	19	45

Final usability problems and severity levels

Table 5.16 presents the number of problems for different severity levels from the three TA conditions. The results show that 31% (19 problems) of the total problems extracted from the CTA method were high severity problems (with critical and major effects). However, for the SC condition, 27% (16 problems) of the final problems were high severity problems, and for the AI condition, 25% (16 problems) were high severity problems. The majority of unique problems identified in each TA condition were at a low level of severity (with minor and enhancement effects), 62% for the CTA condition, 75% for the SC condition, and 63% for the AI condition.

Table 5.16: TA methods and final problem severity levels

	CTA		**SC**		**AI**	
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**
Critical	0	4	0	4	0	4
Major	5	10	3	9	2	10
Minor	10	28	9	31	12	31
Enhancement	1	2	0	2	5	0
Total	16	44	12	46	19	45

Looking at the manner in which the unique problems were detected, the analysis revealed that all the low severity problems in the CTA and SC conditions were from the verbalisation source, whereas 88% of low impact problems in AI condition were verbalised problems (see Table 5.17).

Table 5.17: Sources and severity levels for the unique final problems in the three TA conditions

	CTA			SC			AI		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Critical	0	0	0	0	0	0	0	0	0
Major	1	1	3	0	2	1	1	0	1
Minor	0	10	0	0	8	1	2	10	0
Enhancement	0	1	0	0	0	0	0	5	0
Total	1	12	3	0	10	2	3	15	1

Final usability problem types

The 98 final problems discovered on the tested website in this study were classified by the usability experts into 23 navigational problems, 44 layout problems, 13 content problems, and 18 functional problems. Table 5.18 shows the number of final usability problems by their type. The distribution of problems across the four types were similar in the SC and AI conditions, with fewest being content, next, functionality, then navigational problems, and the greatest number being problems related to the layout. The CTA showed a similar pattern with the exception of functionality problems being the fewest number of problems and the content problems being the second last. Regarding the unique problems, the majority of the unique problems found by the three methods were related to the layout problems.

Table 5.18: TA methods and final problem types

	CTA		SC		AI		**Total**
	Unique	**Overlapping**	**Unique**	**Overlapping**	**Unique**	**Overlapping**	
Navigation	3	15	1	15	3	16	23
Layout	7	18	7	19	9	17	44
Content	5	4	1	3	2	3	13
Functionality	1	7	3	9	5	9	18
Total	16	44	12	46	19	45	98

Figures 5.6, 5.7 and 5.8 depict the final problems detected according to their types and severity level in each TA method. As illustrated in the figures, the four critical problems

found by the three methods were relating to one navigational problem, one layout problem, and two functionality problems.

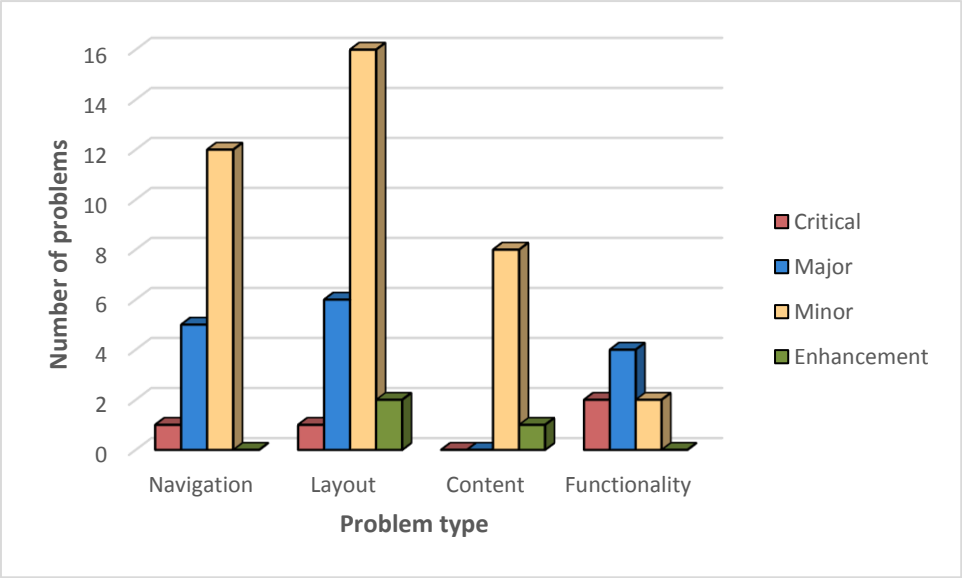


Figure 5.6: Types and severity levels for the final problems in CTA condition

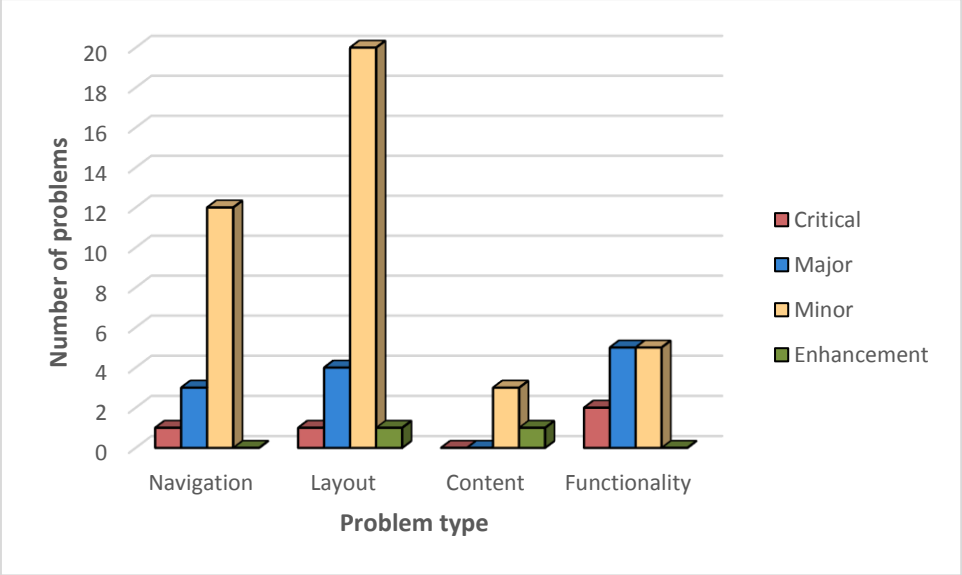


Figure 5.7: Types and severity levels for the final problems in SC condition

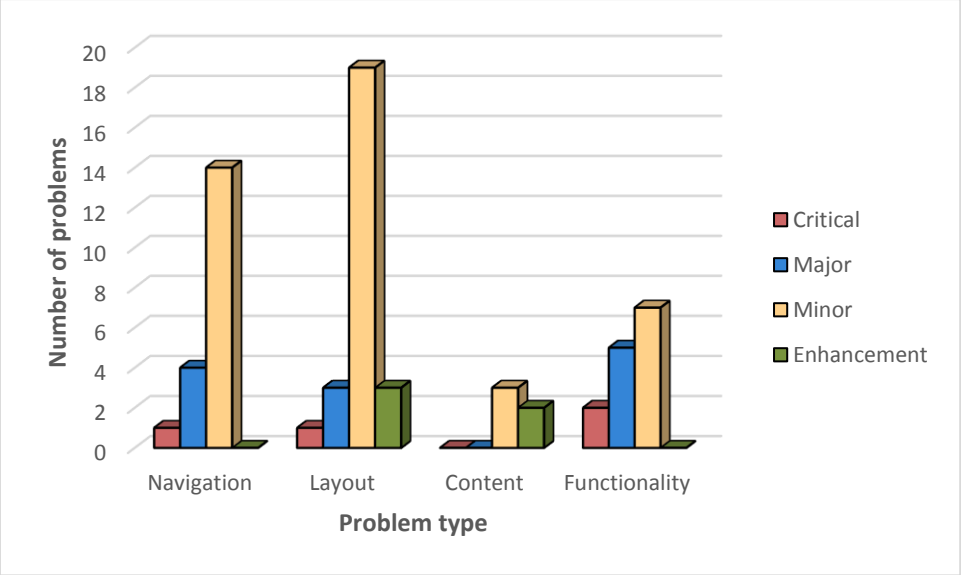


Figure 5.8: Types and severity levels for the final problems in AI condition

An analysis of the unique problems (47 problems) according to their problem sources and types is shown in table 5.19. The results suggest that for the three conditions, all problems relating to content were from the verbalization source.

Table 5.19: Sources and types for the unique final problems in the three TA conditions

	CTA			SC			AI		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Navigation	0	3	0	0	1	0	1	1	1
Layout	1	4	2	0	7	0	0	9	0
Content	0	5	0	0	1	0	0	2	0
Functionality	0	0	1	0	1	2	2	3	0
Total	1	12	3	0	10	2	3	15	1

Looking at the unique problems according to their problem type and severity levels as shown in Table 5.20, results indicate that for AI condition, all the unique problems related to the navigation, layout, and content had low severity. By contrast, for the CTA condition, 66% of the navigational problems and 57% of the layout problems were at low severity level. For the three conditions, all problems relating to content were at low severity level.

Table 5.20: Types and severity levels for the unique final problems in the TA conditions

	CTA				SC				AI			
	Critical	**Major**	**Minor**	**En.***	**Critical**	**Major**	**Minor**	**En.**	**Critical**	**Major**	**Minor**	**En.**
Navigation	0	1	2	0	0	0	1	0	0	0	3	0
Layout	0	3	3	1	0	1	6	0	0	0	6	3
Content	0	0	5	0	0	0	1	0	0	0	0	2
Functionality	0	1	0	0	0	2	1	0	0	2	3	0
Total	0	5	10	1	0	3	9	0	0	2	12	5

*Enhancement

Reliability of problem identification and classification

As in the previous experiment, an additional evaluator was recruited to carry out an inter-coder reliability check on the usability problem analysis. The independent evaluator in this study in the area of usability testing received his PhD under the supervision of Dr Pam Mayhew.

The second evaluator coded the usability problems for the first participant and discussed his disagreements with the researcher. He then independently analysed six randomly selected testing videos (two from each condition). The any-two agreement formula provided by Hertzum and Jacobsen (2001), explained in section 4.9.4.2, was used to calculate the inter-coder reliability across the six videos. The average any-two agreement for the individual problem identification across the six videos was 70% (individual agreements were: 73%, 71%, 69%, 66%, 75%, and 70%). The any-two agreement for the final usability problems was 75% (CTA: 75%, SC: 73%, and AI: 77%), a very good figure.

The reliability of the coding of the problem source and severity level was examined using Cohen's Kappa (Field, 2009). For the individual problem levels, the resulting Kappa value for the problem source was 0.842 and for problem severity it was 0.671. For the final usability problems, the resulting Kappa value for problem source was 0.885, and the severity level was 0.724. This correlates a high reliability for the coding.

Figure 5.9 shows a selection of problems as they occurred in the usability test approaches.

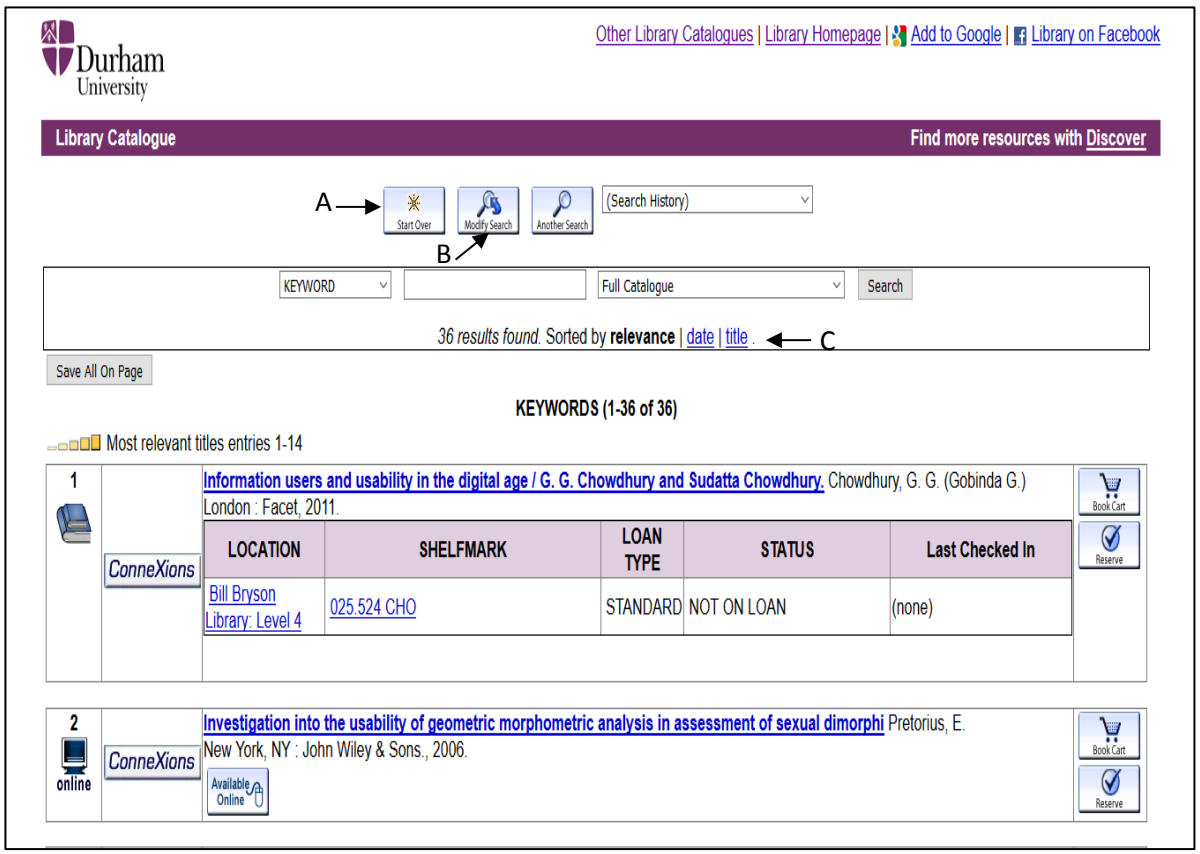


Figure 5.9: Illustration of some usability problems discovered: A) Two confusing buttons in the results page "start over" and "another search"; B); "Modify Search" button is not properly worded. It should be changed to "Advanced Search"; C) There is no option to sort items by publisher.

5.10.5 Comparative Cost

The cost of employing the three TA methods under study was measured by recording the time the evaluator spent conducting firstly testing and latterly analysing the results for each method. As mentioned in section 3.9.4, the session time, recorded via an observation sheet (Appendix D8), refers to the time taken to carry out the entirety of testing sessions, including: instruction of participants, data collection, and time spent solving problems which arose during sessions. The analysis time, monitored throughout via a web-based free-time tracking software called Toggle (Version, 2013), means time taken to extract usability problems from each method's testing datum. The sum amount of time spent on these actions was finally utilised for a comprehensive costing evaluation of the methods. The following sub-sections review the approximate time taken for each TA method (section 5.10.5.1) and provide, using industry standards, an estimation of their financial cost (section 5.10.5.2).

5.10.5.1 Temporal Cost

Table 5.21 depicts time the evaluator (the author) spent applying and analysing the results for the three methods. As is shown in the table, the AI method required the longest session time (844 minutes), whereas the CTA method required the shortest session time (723 minutes). The SC testing lasted for 775 minutes. The total time taken to apply the three verbalization methods was 2342 minutes.

Table 5.21: TA methods and time expense

	CTA	**SC**	**AI**	**Total**
Session time *(m)*	723	775	844	2342
Analysis time *(m)*	865	912	980	2757
Total time *(m)*	1588	1687	1824	5099

One-way ANOVA test was conducted to determine if there were significant differences in the mean session time especially between the RTA and HB conditions. The Shapiro-Wilk test showed that the data were approximately normally distributed for the three TA groups, with p= .087 for the CTA group, p= .492 for the SC group, and p= .513 for the AI group, respectively. The assumption of homogeneity of variances was also met (p= .832). ANOVA testing with a Tukey post-hoc analysis revealed that the session time in the AI was significantly longer than in the CTA condition (see Table 5.22). No significant difference was found between the SC and AI conditions or the CTA and SC conditions.

Table 5.22: Session time for the TA methods

	CTA		**SC**		**AI**		**Value**
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Session time (*m*)	36.00*	6.25	38.50	6.96	42.05*	6.28	F(2,57)=4.12, p=0.021

* AI differed significantly from CTA ($p< 0.05$)

The total amount of video footage of the evaluation sessions was more than 1682 minutes of videos, being 503 minutes of recordings of evaluations by CTA participants, 555 minutes by SC participants and 624 minutes by AI participants. The total time taken to identify usability problems using the three methods was 2757 minutes, with the AI method requiring the highest amount of time (980 minutes) in comparison to the CTA (865 minutes) and SC methods (912 minutes). The Shapiro-Wilk test indicated that the data were not normally distributed for the AI group with p= .018. Accordingly, it was not appropriate to use ANOVA testing, and instead the nonparametric Kruskal Wallis H test with Bonferroni

post-hoc analyses were used and showed that the analysis time in the CTA was significantly shorter than in the AI condition (see Table 5.23). The lengthy time spent on the analysis of AI condition is unsurprising, as prolonged session times will inevitably lead to a longer analysis process.

Table 5.23: Analysis time for the TA methods

	CTA		SC		AI		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Analysis time (m)	43.25*	4.91	45.60	3.36	49.00*	4.83	$\chi 2(2)=8.23$, $p=.016$

* p< 0.05 significance obtained

The overall results showed that the CTA method incurred the shortest time (1588 minutes), followed by the SC method (1687 minutes) and then the AI method (1824 minutes). The total time taken for the conducting of methods and the analysis of the usability problems discovered by three methods was 5099 minutes. By dividing the time the evaluator spent on a method by the number of problems identified by that method, the time needed per problem detection can be computed and compared (Als et al., 2005). The CTA method had an estimated temporal cost of 26.46 minutes per usability problem, whereas the SC and the AI had a cost of 29.08 and 28.50 minutes per usability problem respectively (see Table 5.24).

Table 5.24: TA methods' temporal costs per problem

	Time spent (*m*)	**Problem found**	**Time per Problem (*m*)**
CTA	1588	60	26.46
SC	1687	58	29.08
AI	1824	64	28.50
All	5099	98	52.03

5.10.5.2 Financial cost

Martin et al. (2014) provided information on the daily rate usability evaluators charge for usability consultation at £800.00 per 7.5-hour day. This means that the hourly fee for usability consultation is approximately £107. This figure can be compared to the data from Section 5.10.5.1 to produce the financial costs for the methods if the methods were being conducted in a business environment. Table 5.25 shows the amount of evaluator hours spent conducting and analysing the results of each method times the hourly cost of a usability evaluator. This produces the total financial cost of each TA evaluation (rounded to the nearest pound). It is reported in Table 5.25 that CTA testing would cost £2831, which

is less than the cost of SC testing (£3007) and AI testing (£3252). The cost of the application and analysis of the three methods would be £9093.

Table 5.25: TA methods' financial cost

	Evaluator Minutes	Evaluator Hours	Hourly Fee	Financial Cost
CTA	1588	26.46	£107	£2831
SC	1687	28.11	£107	£3007
AI	1824	30.40	£107	£3252
All	5099	84.98	£107	£9093

By comparing the financial costs of each method against the amount of problems it found, the financial cost per problem can be deduced and compared (Martin et al., 2014) (see Table 5.26). The CTA testing produced the cheapest cost per problem found at £47 compared to the SC and AI methods which cost per problem for both methods found at £51.

Table 5.26: TA methods' financial costs per problem

	Financial Cost	Problem found	Cost per Problem
CTA	£2831	60	£47
SC	£3007	58	£51
AI	£3252	64	£51
All	£9093	98	£93

The overall picture created in this section is that the CTA is a more cost-effective method than SC and AI testing.

5.10.6 Relationship between Sample Size and Number of Problems Detected

One of the questions this research sought to address is whether the relationship between the sample size and the number of problems detected work differently for the TA methods under investigation. As mention in Section 2.4.1, it has been argued by Nielsen (2000) that five test participants are enough to find 85% of usability issues. The first experiment did not achieve the results this magic number promises. In this study, as reported in section 5.10.1.2, the three TA groups produced 98 usability problems in DU-L website, of which 85% would be 84 problems. Despite all groups using twenty test participants, which is four

times the recommended number, none of the groups generated this many problems (see Section 5.10.1.2). This confirms that the 'five participants' argument is at the very least controversial in usability testing. Nevertheless, the proportion of issues detected by five participants from each of the TA method was examined to determine whether or not the methods show similar patterns. This section starts by exploring the number of problems discovered by the best and first participants from each TA condition (5.10.6.1). It then determines the number of participants needed to find 85% of problems for the whole test and for each condition (5.10.6.1).

5.10.6.1 Number of Problems Discovered by the Best and First Five Participants

Table 5.27 reports the performance of the top five participants in each TA condition. The T-CTA, T-SC, and T-AI consist of the top (T) performing five participants who discovered the most problems for the CTA group (T-CTA), SC group (T-SC), and AI group (T-AI), respectively. As evident in Table 5.27, the T-CTA, T-SC, and T-AI groups uncovered only 25%, 23%, 28% respectively of the final number of problems found on the tested website which is significantly less than the 85% claimed by Nielsen (2000), concurring with the results found in Study One.

Table 5.27: Top (T) five participants and number of problems discovered (absolute and percentage of total number)

Top performing five participants								**(Nielsen, 2000)**		**Maximum to be discovered**	
T-CTA		**T-SC**		**T-AI**		**All groups**					
#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
25	25%	23	23%	28	28%	40	40%	84	85%	98	100%

Figure 5.10 depicts the overall relationship between the sample size and number of problems discovered in each TA condition. As shown in the figure, the first five participants from the CTA, SC, and AI groups were only able to uncover 18%, 20%, and 23% respectively of the final usability problems detected in the DU-L website. The first ten participants manged to detect 31% of the problems in the CTA condition, 28% in the SC condition, and 35% in AI condition. The number of usability problems found increased with the addition of each new participant until the nineteenth participant in AI condition. Generally, it can be said that the relationship between sample size and percentage of problems detected for the three were very similar.

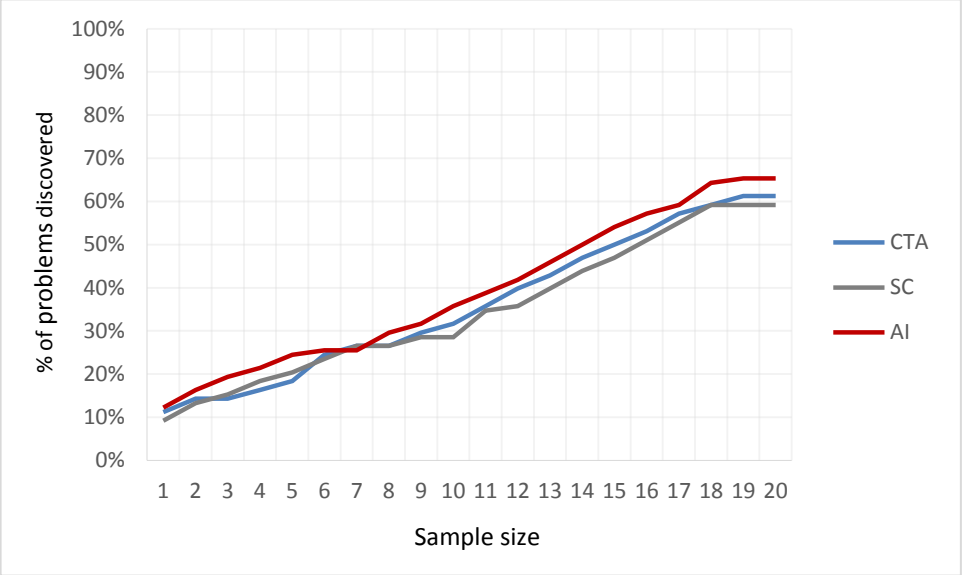


Figure 5.10: All participants' performances in the three TA conditions (cumulative)

The analysis in the following section specifies the sample size required to find 85% of the problems.

4.10.6.2 The Sample Size Required to Detect 85% of Problems

As explained and applied in section 4.9.6.2, the average detection rate of usability problems must be first computed in order to be able to calculate the sample size needed to detect a pre-set percentage of problems. In this study, the detection rate is 0.084, which means that 26 test participants would be needed from the whole sample participated in the experiment (60 participants) to detect 85% of the final number of usability discovered by the three methods (98 problems). The following table (5.28) shows the number of participants needed and consequently achievable percentages of finding usability problems.

Table 5.28: Participant number and the targeted percentage of problems

Targeted Parentage	Sample size required
99%	57
95%	45
90%	39
85%	26
75%	22
50%	18

Looking at the number of test participants required by each TA method in order to detect 85% of the number of the problems found, the adjusted average detection rate (explained in section 4.9.6.2) of usability problems was 0.055 in the CTA, 0.051 in the SC, and 0.059 in the AI, so the sample size needed to reveal 85% of the problems is 33 participants for the CTA method, 35 participants for the SC method, and 31 participants for the AI method.

5.10.7 Correlational Analysis of Usability Measures

The usability measures analysed in the preceding sections may correlate. Similar to the previous study, the correlations between the most common usability measures were analysed. This comprised the final number of usability problems detected, task success rate, time on task, participants' satisfaction with the website (i.e., SUS score), mouse clicks, and browsed pages. The Spearman's correlation coefficient (explained in section 4.9.7) was used to investigate whether or not there are associations between the variables.

Table 5.29 summarizes the correlation analysis for the three TA methods. Across all six measures, the correlations for classic and relaxed thinking aloud were very similar. This table shows the following main results:

- There is a strong, statistically significant, positive relationship between time spent on tasks, mouse clicks and visited pages. This result implies that the participants who spent more time, made more mouse clicks and visited more pages.
- There is statistically significant positive relationship between browsed pages and mouse clicks.
- There is no statistically significant relationship between time spent on tasks and the number of usability problems discovered in the three TA groups. However, it should be noted that the relationship was almost significant.
- There is no statistically significant relationship between problems discovered and participant satisfaction with the website in the TA conditions.
- There is no statistically significant relationship between task performance measures and participant satisfaction.

Table 5.29: Correlations amongst usability measures (N=20)

Usability measures		Task success	Task time	SUS score	Browsed pages	Mouse clicks	Usability problems
Task success	CTA	1	.175	.254	.128	.132	-.128
	SC	1	.254	.178	.195	.187	-.142
	AI	1	.077	.295	.103	.110	-.291
Task time	CTA		1	-.239	.715*	.802*	.443
	SC		1	-.414	.708*	.835*	.461
	AI		1	-.423	.823*	.864*	.485
SUS score	CTA			1	-.178	-.258	-.354
	SC			1	-.064	-.164	-.189
	AI			1	-.255	-.172	-.398
Browsed pages	CTA				1	.586*	.354
	SC				1	.605*	.337
	AI				1	.639*	.389
Mouse click	CTA					1	.246
	SC					1	.315
	AI					1	.349
Usability problems	CTA						1
	SC						1
	AI						1

* Correlation is significant at the .0001 level (2-tailed)

5.11 Discussion

This empirical study has focused on the consequences of using two relaxed think-aloud protocols on the utility and validity of the usability data collected. Below, the results obtained from this study are compared to some of the related work and the implications for usability evaluation are discussed. Table 5.30 offers an overview of the main findings of the present study.

5.11.1 Think-Aloud Methods and Participants' Task Performance

As shown in section 5.10.2, there were significance differences between the AI condition and the other two conditions in the participants' task performance. The use of proactive interventions in the AI condition slowed down the process of task solving and led to a higher number of mouse clicks and pages viewed compared to the CTA and SC conditions.

Ericsson and Simon (1993) warned that the practitioners' use of interventions could disrupt participants thought process, causing a change in this process and task performance or what they referred to as "reactivity". This implies that the significant increase in task time and navigational behaviour is due to the triple-workload effect of the AI condition, in that participants needed to solve the task, think aloud, and also respond to the evaluator's questions. This finding, therefore, casts doubt on using task outcome in an AI evaluation as an overall indication of the usability of an artefact, and on the implicit assumption that the problems found in an AI usability test are by definition real user problems. These results were in line with Hertzum et al. (2009). However, the findings contradicted Olmsted-Hawala et al. (2010) who found that the evaluator's probing improved participants' task solving accuracy. One explanation may be that the two studies mentioned above did not take steps to control the participants' individual differences by matching them as closely as possible between conditions, and also used different evaluators between different conditions. These additional variables may affect the results.

The SC participants performed their tasks neither better nor worse than the participants in the CTA condition. This corresponds to earlier findings by Olmsted-Hawala et al. (2010). This finding implies that practitioners have a choice between using the traditional TA mode put forth by Ericsson and Simon (1993) or the newer mode suggested by Boren and Ramey (2000), as these two conditions show no statistically significant differences in task solving accuracy, efficiency or navigational behaviour.

5.11.2 Think-Aloud Methods and Participants' Experience

For the participants' satisfaction with the tested website, although the AI condition increased participants' task completion time and changed their navigational behaviour, it did not lead to changes in their perceptions about the usability of the websites compared to the classic and SC conditions. This finding is in disagreement with the findings of Olmsted-Hawala et al. (2010) who found that participants in the AI were significantly more satisfied with the website compared to participants in CTA and SC. This conflicting result may be explained by the inevitable differences in experimental design, task set and interface. Another plausible reason could be the low correlation existent in this study between task performance and the participants' satisfaction which was also proved in

Study One and numerous other studies (Hornbæk and Law, 2007; Nielsen and Levy, 1994b).

With regard to the participants' experience with the TA testing, the evaluator seems to have had a more detrimental effect in the AI condition than in the CTA and SC conditions, with participants indicating the presence of the evaluator as a disturbance. AI participants also felt that their working procedure on the tasks were significantly slower than their CTA and SC counterparts. Once again, as mentioned earlier, these discrepancies can perhaps be explained by the evaluator's probing.

The CTA and SC participants in the current study appeared to have similar testing experiences. Most measures of experience with the TA test questionnaire yielded neutral to positive judgments for the two evaluation methods. Accordingly, it can be concluded that the ecological validity of these two methods (i.e. a method should be comfortable for participants to use) is ensured. No previous study has investigated the participants experience with relaxed TA methods, so no comparison can be made. In summary, the findings would seem to suggest that given the choice, participants would prefer to use the CTA or the SC methods rather than the AI method.

5.11.3 Think-Aloud Methods and Usability Problems Identified

Contrary to general emphases on the AI protocol, this study showed no indication that it was superior for identifying usability problems. At the individual problem level, the three conditions yielded a similar number of problems, and no differences were found in terms of problem source. The AI method only identified a higher number of problems with enhancement effect than the CTA and SC conditions. Considering the problem types, the CTA identified a higher number of content problems than the SC methods. However, both the difference in problem severity and types concern a small proportion of problems. At the final problem level, the AI method enabled the detection of only four more final problems. This was at the cost of putting the ecological validity of the method under threat, and the likelihood of false problems. In contrast, the SC method produced slightly fewer issues than the CTA method. In all, the overall picture that arises is one in which the three methods are comparable in terms of number and types of problems detected. As stated in the above section, no existing study has examined the impact of relaxed methods on the

quantity and quality of usability problems so the results of this study cannot be compared with the literature.

5.11.4 Think-Aloud Methods and Cost

The findings of this study reveal that the CTA method cost less in comparison to the SC method and significantly less in contrast to AI method in terms of the total time required by the evaluator to conduct the testing and identify the usability problems. Moreover, the financial cost of the CTA method was estimated to be less than the other two methods. In comparison to Study One, the CTA was slightly more expensive in this study; this may be attributed to the higher number of the tasks in this study, which prolonged the time of the test session and the analysis process. No previous studies have compared the cost of employing relaxed TA variations, so no comparison can be made.

5.11.5 Think-Aloud Methods and Sample Size Needed

Having investigated the relationship between the sample size and the number of problems identified by the TA conditions in detail, two conclusions can be drawn. First, the controversial argument that five participants is enough to identify 85% of problems was not verified here. The results for the best performing five participants from the three conditions did not find 40% of total problems discovered. Furthermore, the performance of the first five participants from the three conditions did not exceed 23%. These findings are in agreement with Study One, and other studies supporting the argument that five users are not enough (Molich et al., 2004; Lindgaard and Chattratichart, 2007). The second conclusion is that the relationship between sample size and percentage of problems detected for the three showed similar patterns.

Table 5.30: Overview of the main findings of the relaxed think-aloud study

Results in terms of	The relaxed TA study
Task performance	
- Successful task completion	No difference between the three TA methods
- Task duration	AI participants spent more time on tasks than CTA and SC participants
- Mouse clicks	AI participants clicked their mouse more than the CTA and SC participants
- Browed pages	AI participants visited more pages than the CTA and SC participants
Participant experiences	
- The tested website	No difference between the three TA methods
- The TA method	AI participants felt they worked slower and were more distracted by the evaluator than CTA and SC participants
Usability problems	
- Individual problems	
Detection means	No difference between the three TA methods
Source of problems	No difference between the three TA methods
Severity of problems	AI produced higher number of enhancement problems than CTA
Types of problems	CTA produced higher number of content problems than SC
- Final problems	
Detection means	No difference between the three TA methods
Source of problems	No difference between the three TA methods
Severity of problems	No difference between the three TA methods
Types of problems	No difference between the three TA methods
Unique problems	CTA: 16, SC: 12, AI: 19
Methods Cost	
- Temporal cost	CTA required less time than the SC and AI methods
- Financial cost	CTA would require less financial cost than the SC and AI methods
Sample size needed	No difference between the three TA methods

5.12 Summary

This chapter has compared the performance of the traditional concurrent think-aloud method with two interactive versions of the method: the active intervention and the speech-communication methods. The three methods were compared through an evaluation of a library website, which involved five points of comparison: overall task performance, test participants' experiences, quantity and quality of usability problems discovered, the cost of employing methods, and the relationship between sample size and number of problems detected.

The study showed that the evaluator's active interventions modified participants' behaviour at the interface and affected negatively their feelings towards evaluation. The three protocols facilitate identification of a similar number of usability problems and types. The traditional protocol generated more usability problems in the content category than

the speech-communication, and the active interventions produced more enhancement problems. However, both of these differences concern a small proportion of problems. The AI method required considerably more time on the evaluator's part and is therefore liable to cost financially more than the other two methods. Lastly, the three methods showed similar patterns in the relationship between sample size and the number of problems discovered.

Although the traditional and speech-communication methods provided similar results in this study to a large extent, the former method enjoys one critical advantage over the latter: directness and simplicity of application. The simplicity of Ericsson and Simon's (1993) classic technique means that it can be applied consistently, whereas the effectiveness of evaluator interaction with participants in the speech-communication protocol is a variant, related to the evaluator's own skills and personal characteristics (Boren and Ramey, 2000). Also, the evaluator's tones of voice, attitude, and friendliness may affect participants' subsequent verbalisations (Rubin and Chisnell, 2008). These actualities, besides the findings of this study - which showed no marked benefit for additional interaction in the speech-communication – allied with particular negative effects of the evaluator's active interventions, suggest it is wiser, safer and cheaper to follow Ericsson and Simon's (1993) concurrent classic think-aloud.

The next study will investigate the benefit of adding an additional participant to the test session by comparing performance of the classic concurrent think-aloud method with co-participation method, wherein a pair of participants work together to perform their tasks and engage in verbalizing as they interact.

6

CO-PARTICIPATION STUDY

6.1 Overview

The previous chapter compared the performance of Ericsson and Simon's (1993) classic concurrent think-aloud methods with that of two relaxed versions of the method, namely the active intervention protocol and the speech communication protocol. The results suggested that the concurrent think-aloud method was more efficient in collecting valid usability data.

Another increasingly common variation of the think-aloud method is the co-participation method—also known as the team think-aloud or constructive interaction method—which, in contrast to the other methods, involves two participants per test session. Two people work together to perform their tasks, and verbalise their processes as they interact with the interface and with each other. This chapter presents the third and final empirical study of this research, which explores the use of the co-participation method within website usability testing. The chapter starts by stating the motivations behind the study, defining its specific aims, identifying the test object and tasks, and outlining the participant recruitment procedure. After an overview of the experimental procedure, the chapter then presents the results of the pilot and main experiments. Finally, the chapter concludes by discussing and summarizing the results of the study.

6.2 Motivations

This study was partly inspired by Nielsen (1993a), who recommends enhancing the ecological validity (i.e. the extent to which to a method is comfortable for participants to use) of a test situation by having participants interact, not with the test evaluator, but with a second participant. The number of extant studies on co-participation in the context of website usability evaluation is limited. Adebesin et al. (2009), Als et al. (2005), and Van den Haak et al. (2004) have compared the utility of the co-participation method with single-participant methods of website usability testing. Their studies, however, have a serious common drawback in that they failed to control for the "evaluator effect" on the usability problem extraction process, a factor that might have significant negative consequences on the validity of the comparative study, as explained in section 2.5.3. In addition, Adebesin et al. (2009) did not report on the number and kinds of problems detected by the participants in the think-aloud conditions. With problem detection typically being one of the most important functions of usability testing, the researchers thus failed to account for

a crucial factor in their comparison of the two methods. Furthermore, in Van den Haak et al's (2004) study, another important issue was not taken into account: the level of acquaintance between the pairs. Previous studies have indicated that test participants can behave quite differently depending on how well they know each other (Als et al., 2005). These variables, if not accounted for, can make it difficult to determine cause and effect. The usefulness of the co-participation method is therefore yet to be examined in detail.

6.3 Study Aims

This study aimed to compare Ericsson and Simon's (1993) traditional concurrent think-aloud (CTA) protocol with the co-participation (CP) method in order to determine the benefit of adding an additional participant to the testing session. The two methods were compared through an evaluation of a library website, and their relative validity and utility were measured using five points of comparison: overall task performance, test participants' experiences, quantity and quality of problems discovered, the cost of employing each method, and the relationship between the sample size and the number of problems detected in each test condition.

6.4 Test Object and Tasks

The website (i.e., the Durham University library website) and task set used in the previous study (see sections 5.4 and 5.5) were the same ones targeted in this study. There were a number of factors supporting this decision. Firstly, this study is directly linked to one of the previous experiment's conditions (CTA condition). Secondly, there had been no changes to the website design; the author made another inspection to confirm that the identified problems were still present in the website, and contacted the administrator to confirm that there were to be no modifications in the website's design for the whole duration of the study. Thirdly, the time between these two experiments was short: it did not exceed three months.

6.5 Participants

The recruitment criteria for this study were the same as the ones applied in the previous study (see section 5.6). The sample was recruited through various channels, such as personal emails, posters displayed on schools' notice boards, requests on social networking sites, and conversations with personal contacts. In addition, an email was also sent through official channels to students studying in the researcher's university. The email informed prospective participants that they would be asked to invite a friend to join them in the test session, and that they and their friend would each receive £5 as a token of appreciation for participating in the study. The email also provided a link to the online pre-experiment questionnaires, where prospective participants could provide key demographic details about themselves.

Twenty students who met the study requirements were invited via email to participate in the study. An attempt was made to recruit participants with similar characteristics to the participants in the previous CTA study to mitigate the impact of individual differences. The invited participants were then asked to bring a partner to join them in the session, making a total of forty participants, divided into small teams of two. The students were informed that their partners should have, to some extent, similar characteristics to them in terms of gender, age, Internet experience, etc. The students were also asked to direct their partners to fill out the pre-experiment questionnaires. This method of sampling is known as snowball sampling, and is quite effective in generating a large number of participants with minimal effort (Creswell, 2009).

6.6 Experimental Procedure

All the CP experimental sessions were held in the same laboratory in the school of Computing Sciences at UEA. Permission to run the study was sought and granted from the University's Ethics committee (see Appendix E1). The experimental procedure in the CP condition was as follows[25]. Upon arriving, the evaluator (the author) welcomed the participants to the laboratory and made them feel at ease, after which they were informed that they were going to be evaluating a library website. Next, every participant was asked to read and sign a consent form (see Appendix D5). After signing the consent forms, the

[25] For the CTA experimental procedure see section 5.8 in Chapter 5

paired participants were seated at the computer—one of them sitting in front of it, and the other next to it—and were given a maximum of two minutes to familiarise themselves with the lab computer. Before beginning the task, they were explicitly instructed to work together, in these words: "even though only one of you can actually control the mouse, you have to perform the tasks as a team by consulting each other and making joint decisions. I also want you to state aloud what you are doing". They were also told not to turn to the evaluator for assistance (see Appendix E2). Participants subsequently engaged in a brief practice session using the simple, neutral task of looking up the word '*chant'* in an online dictionary. On completion of this step, the participants then began the experiment proper. During the testing sessions, the evaluator remained in the same room as the participants, and only issued think-aloud reminders if the participants fell silent for 15 seconds. The Morae software (2015) was used to record the computer screens and participants' voices. Once the participant pairs had completed the tasks, each individual participant was asked to fill in, without collaboration, the two online post-test questionnaires to provide feedback on the evaluated website (the System Usability Scale questionnaire, see Appendix B2) and the testing experience (Experience with TA Test questionnaire, see Appendix B1). Finally, the evaluator thanked[26] participants for taking part, and gave each one of them the promised £5 as a token of appreciation for participating in the study.

Figure 6.1: CP condition (picture taken with participants' permission)

6.7 Results

This section presents the following results of both the classic CTA and the CP conditions: 1) participants' task performance; 2) participants' experience of the test; 3) quantity and

[26] https://www.techsmith.com/morae.html

quality of problems that were collected; 4) the cost of employing each method; 5) the relationship between the sample size and the number of problems detected; and 6) a correlational analysis of the usability measures used.

6.7.1 Participants' Profiles

Table 6.1 illustrates the summary statistics of the demographic characteristics of the CP participants in the present study. These are presented alongside the details of the participants from the previous CTA study. As mentioned before, an attempt was made to recruit participants with similar characteristics to the participants in the previous CTA study.

The participants in the CP condition were working in pairs, each with a different role. The "CP actor" column in Table 6.1 refers to the participants working behind the computer in the CP condition, while the "CP co-actor" column refers to those sitting next to the CP actor. As shown in Table 6.1, 24 men (60%) and 16 women (40%) participated in the CP experiment. 60% of the CP participants were aged between 18 and 29, 35% between 30 and 39, and 5% between 40 and 50. All participants were frequent users of the Internet, and had not visited the targeted site prior to this study. The author believes that the independent participant groups were matched successfully, given that a non-parametric Kruskal-Wallis H test (Kruskal and Wallis, 1952) found no statistically significance difference between the think-aloud conditions in terms of nationality ($\chi^2(2)= 0.606$, $p= .739$), gender ($\chi^2(2)= .555$, $p= .758$), age ($\chi^2(2)= 1.78$, $p= .411$), or Internet use ($\chi^2(2)= .284$, $p= .241$). Accordingly, it can be stated that the internal validity of the study is ensured.

Table 6.1: Summary statistics of demographic characteristics of participants

Characteristics		**CTA** (n=20)	**CP actor** (n=20)	**CP co-actor** (n=20)	**Total** (n=60)	**Percent**
Country	British	15	13	13	41	68.33
	European	5	7	7	19	31.66
Gender	Male	13	13	11	39	65
	Female	7	7	9	21	35
Age	18-29	11	14	10	35	58.33
	30-39	9	4	10	23	38.33
	40-50	0	2	0	2	0.033
Internet use	Daily	18	14	17	49	82
	At least once a week	2	6	3	11	18

6.7.2 Task Performance

Four indicators were used in this study to measure the task performance in the CP condition and determine whether the method induces reactivity (i.e. a change in task performance caused by the parameters of the task). These indicators included: the number of tasks that were completed successfully, the total amount of time required to complete the tasks, the number of mouse clicks made, and the number of pages visited. The following subsections show the task performance of the CP participants, and how their behaviour compared with their CTA counterparts. The CTA condition was regarded as the control group in this study, as it was shown in Study One and Study Two that classic CTA has no effect on task performance.

6.7.2.1 Task Completion

The task completion metric was used to determine whether the CP group were able to successfully complete more tasks than the CTA group. The average completion rate for each participant over the nine tasks was calculated. Table 6.2 shows the completion rate of both of the groups. Participants in the CP condition can be seen to have achieved a higher success rate (65%) than participants in the CTA condition (who achieved a 61% success rate). In other words, each participant pair in the CP group completed an average of 5.85 out of the nine tasks, whereas participants in the CTA group completed an average of 5.50 tasks each.

Table 6.2: Descriptive statistics of the task completion rates for the TA methods

Task completion	**CTA**	**CP**
Total number of tasks	180	180
# of successful tasks	110	117
Percentage of successful tasks	61%	65%

To determine whether this difference in averages is significant, an independent t-test was conducted. The independent t-test is a parametric test used to compare the means of two unrelated groups, and assumes the approximate normal distribution of the data, and the homogeneity of variances, though violation of the latter assumption tends not to be a serious issue if the sample size in each group is similar (Filed, 2005). As mentioned in previous chapters, for data distribution to qualify as approximately normal, the p-value of the Shapiro-Wilk test must be more than 0.05 for each group of the independent variable. To meet the assumption of homogeneity of variance, the *p*-value of the Levene's test must be more than 0.05.

Task success rates were approximately normally distributed for the two test groups, as verified by the Shapiro-Wilk test, with a result of p=.076 for the CTA group, as mentioned in Chapter Five, and p=.326 for the CP group. The second assumption of the t-test was violated (p=.017). Accordingly, an independent t-test test based on equal variances not assumed was run, and revealed no significant difference in the number of successful task completions between the two conditions, as shown in Table 6.3.

Table 6.3: Inferential statistics of the task completion and the TA methods

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Task success	5.50	1.09	5.85	1.79	t=.-1.72, df= 38, p= .096

6.7.2.2 Time on Task

The time-on-task metric measured the time taken by participants to complete each individual task, and also the time taken to complete all nine tasks. This metric looked solely at task time, regardless of whether the tasks were completed successfully. Table 6.4 shows the total time spent by all participants on the nine tasks, and the mean time spent by each participant or pair on all nine tasks.

Table 6.4: Descriptive statistics of time on tasks for the TA methods

Time on task	CTA	CP
Overall time spent on tasks (m)	503	562
Mean time spent on tasks (m)	25.15	28.10

Examining these results reveals that the participants in the CP condition took longer to complete the tasks compared to the participants in the CTA condition. The CP group spent a total of 562 minutes on tasks, whereas the CTA spent a total of 503 minutes. An independent t-test was conducted in order to determine if there were statistically significant differences in the mean time spent on all tasks. The Shapiro-Wilk test showed that the task times were approximately normally distributed for the two think-aloud groups, with p= .099 for the CTA group, and p= 0.086 for the CP group. The assumption of homogeneity of variances was not met because the p-value of Levene's test was less than 0.05 (p= 0.001). The results of an independent t-test based on equal variances not assumed indicated that there is no significance difference in the time-on-task performance metric (see Table 6.5).

Table 6.5: Inferential statistics of time on tasks and the TA methods

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Time on tasks (*min*)	25.15	3.45	28.10	5.70	t=.-1.74, df= 38, *p*= .093

6.7.2.3 Navigational Behaviour

In this study, the MORAE software was also used to explore the navigational behaviour of the CP participants through collecting data such as mouse clicks and the number of pages visited (for more details on MORAE software see section 5.7). To determine whether there is a significant difference in navigational behaviour between the test conditions, an independent t-test was conducted after meeting the assumptions of the test. Table 6.6 shows the results of the normality test and the homogeneity of variance test for the navigational behaviour data.

Table 6.6: Tests for normality and homogeneity of variance for the navigational measures

	Shapiro-Wilk test		Levene's test
	CTA	**CP**	
Mouse clicks	.638	.714	.432
Browsed pages	.371	.968	.865

The independent t-test test found no statistically significant difference between the test groups in the number of mouse clicks or pages visited (see Table 6.7). Therefore, the results in this section all suggest that the CP method does not affect participants' task performance; in other words the think-aloud process did not induce reactivity. The next section will discuss the testing experiences of the participants.

Table 6.7: Navigational measures for the TA methods

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Mouse clicks	105.20	22.70	110.60	15.69	t=.-1.53, df= 38, *p*= .134
Browsed pages	34.80	7.86	39.40	11.03	t=.-1.09, df= 38, *p*= .280

6.7.3 Participants' Experiences

As mentioned earlier, the researcher gathered data on the participants' satisfaction with the usability of the test website, as well as their experiences of participating in the test, using two post-test questionnaires: the System Usability Scale (SUS) questionnaire (Appendix B32) and the Experience with TA Test questionnaire (Appendix B1). As the participants

in the CP condition were working in pairs, each with a different role (actor/collaborator) that may have influenced their experiences, they will be treated as separate subgroups in the analyses of the post-test questionnaire results. The actors, i.e. the participants working behind the computer, will be referred to as "CP actor", while the collaborator, i.e. those sitting next to the person working behind the computer, will be referred to as CP co-actors.

6.7.3.1 Participants' Satisfaction with the Usability of the Website

Table 6.8 shows that the participants in the conditions did not find the system usable. Both the actor participants and the co-actor participants in the CP condition gave less SUS score than the CTA participants, which means they were less satisfied with the site. Having met the assumptions of normality (p= 0.448 for the CTA group; p= 0.303 for the CP actor group; and p= 0.082 for the CP co-actor group) and homogeneity of variances (p= 0.254), a one-way ANOVA test was conducted, and indicated that the satisfaction rating did not differ significantly between the conditions.

Table 6.8: Participants' satisfaction with the usability of the tested website

	CTA		CP actor		CP co-actor		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
SUS score	61.60	10.58	54.35	10.90	57.20	7.72	F(2,57)=2.73, p=.073

On a totaled scale of 1 to 100

6.7.3.2 Participant Experience with the TA Test

This section discusses the results of the second post-test questionnaire (the Experience With TA Test questionnaire), which aimed to understand the participants' experiences with (1) how the TA method affected their ability to work normally on the test tasks; (2) having to think aloud concurrently whilst working together; and (3) the presence of the evaluator during testing. As in the previous section, the CP actors and co-actors will be considered separately. A non-parametric test, the Kruskal Wallis H test, was used to analyse the data. This method was chosen because the data were of ordinal nature and were not normally distributed (see Appendix E4). Table 6.9 presents the results of participants' ratings in the TA conditions.

To start with, all participants were asked to assess in what respect(s) their working process during the test differed from their normal working process by estimating how much slower

and more focused they felt they were whilst working on the tasks. As shown in Table 6.10, the participants in all the conditions felt that their work on tasks was not that different from their normal work. The scores for the two items are fairly neutral, ranking around the middle of the scale, with average scores ranging from 2.10 to 3.00. No significant differences were found between the conditions.

Participants were next asked to indicate whether, and to what extent, they felt that having to think aloud and/or work together was difficult, unpleasant, tiring, unnatural, and time consuming. A Kruskal Wallis H test and Bonferroni post hoc analyses showed that both the CP actor participants and the CP co-actor participants found working together significantly more natural and pleasant than the participants in the CTA condition did about having to think aloud concurrently (see Table 6.9). It might be easy to see why working together would be evaluated more positively by participants: participants can share their workload and they can talk to each other in a much more natural way than if they were required to think aloud concurrently whilst working alone.

The final part of the questionnaire concerned the presence of the evaluator. Participants were asked to indicate to what degree they found it unpleasant, unnatural and disturbing to have the evaluator present during the study. Interestingly enough, a Kruskal-Wallis H test and Bonferroni post hoc analyses revealed that the CP co-actor participants found the presence of the evaluator to be significantly more unnatural than did the CTA participants. No such differences arose in other aspects (see Table 6.9). A possible explanation could be the workload of the participants. The CTA participants and the CP actors had to actively perform tasks and think aloud, which considerably reduced the amount of attention they could spare for noticing the evaluator. The CP co-actor participants, on the other hand, were only helping their partners perform tasks, which might require less concentration and thus make them more aware of the evaluator's presence.

Table 6.9: Participants' experience with the TA test

	CTA		CP actor		CP co-actor		Value
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Working condition							
Slower than my normal working	2.50	1.19	3.00	1.02	2.80	1.25	$\chi2(2)=2.31$, $p=.315$
More focused than my normal working	2.70	1.36	2.10	0.85	2.30	0.97	$\chi2(2)=1.17$, $p=.551$
Think-aloud/ Working together							
Difficult	2.10	1.07	1.95	0.75	1.70	0.50	$\chi2(2)=.915$, $p=.633$
Unnatural**	2.85	0.44	2.05	0.64	1.90	0.85	$\chi2(2)=12.45$, $p=.002$
Unpleasant**	2.45	1.14	1.55	0.51	1.30	0.59	$\chi2(2)=14.40$, $p=.001$
Tiring	2.20	1.00	1.80	0.76	1.60	0.52	$\chi2(2)=4.32$, $p=.115$
Time-consuming	2.60	1.45	2.45	0.88	2.30	1.09	$\chi2(2)=1.62$, $p=.922$
Evaluator presence							
Unnatural	1.50*	0.93	1.90	0.71	2.05*	0.68	$\chi2(2)=7.14$, $p=.028$
Disturbing	1.45	1.17	1.40	0.54	1.20	0.32	$\chi2(2)=1.78$, $p=.410$
Unpleasant	1.25	1.23	1.20	0.39	1.50	0.51	$\chi2(2)=4.69$, $p=.096$

Five-points scale (1: Strongly disagree to 5: Strongly agree) * p< 0.05 significance obtained, **p< 0.005 significance obtained

6.7.4 Usability Problems

The aim of usability evaluation is to detect as many usability problems as possible. Therefore, if the quantity and quality of usability problems identified differs between methods, then this important factor should be taken into account when selecting an evaluation method. This section compares the CTA and CP methods in terms of the number and quality of individual (i.e., problems detected per participant/pair) and final usability problems (i.e., problems detected in each condition) that were extracted from the test sessions. Statistical comparisons made at the individual problem level used the non-parametric Mann-Whitney test (Field, 2009), as the data were not normally distributed (see Appendix E5).

6.7.4.1 Individual usability problems

Table 6.10 presents the number of problems discovered during interaction with the website by each testing method, and also categorises all problems according to the way in which they came to light: (1) by observation (i.e., problems detected from observed evidence with no accompanying verbal data), (2) by verbalisation (i.e., problems detected from verbal data with no accompanying behavioural evidence), or 3) by a combination of observation and verbalisation.

A Mann-Whitney test revealed that the CP method detected significantly more individual problems than did the CTA (see Table 6.10). One explanation for this could be the fact that the CP condition had two pairs of eye which might allow them to notice more problems on the interface. Another explanation could be that as the CP condition involves two people, they could both suggest possible ways of carrying out the nine tasks. This collaborative way of working might thus offer more opportunities for the participants to encounter and articulate usability problems. With respect to the manner in which the individual problems were detected, as can be seen from Table 6.10, a Mann-Whitney test reveals that the CP method detected significantly higher number of individual problems through a combination of observation and verbalization.

Table 6.10: TA methods and the number of individual problems

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Observed	2.50	2.06	2.35	1.51	U= 185.5, *z*= -.401 , *p*= .698
Verbalised	2.20	1.28	1.45	0.76	U= 146.5, *z*= - 1.51, *p*= .149
Both*	6.60	3.78	10.90	4.37	U= 311, *z*= 3.01, *p*= .002
Total*	11.30	3.96	14.70	4.61	U= 290.5, *z*= 2.46 , *p*= .013

* p< 0.05 significance obtained

Individual usability problems and severity levels

The individual problems detected were categorised into four types according to their impact on participants' task performance: 1) critical, 2) major, 3) minor, and 4) enhancement (for problem severity coding details see section 4.9.4.1). A Mann-Whitney test found a significant difference between the CTA and CP methods regarding the number of individual problems whose severity was rated as “minor” or “enhancement”. The CP method produced significantly more individual minor and enhancement level problems than did the CTA method (see table 6.11).

Table 6.11: TA methods and individual problem severity levels

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Critical	3.50	0.94	3.25	1.43	U= 151.5, *z*= -1.35, *p*= .192
Major	4.20	1.50	4.55	2.67	U= 194.5, *z*= - .150, *p*= .833
Minor*	3.35	2.45	5.60	2.85	U= 290.5, *z*= 2.48 , *p*= .013
Enhancement*	0.25	0.55	1.30	0.97	U= 321.5, *z*= 3.59 , *p*= .001

* p< 0.05 significance obtained

Individual usability problem types

To enable an examination of the types of problems that were discovered in the CP condition, two usability experts classified all detected problems into four specific problem types: navigation, layout, content, and functionality (for problem type coding details see section 4.9.4.1). The inter-coder reliability was computed using Cohen's kappa (explained in section 4.9.4.1). The overall kappa was 0.94, which shows a highly satisfactory level of inter-coder agreement.

Table 6.12 shows the number of different types of individual problems identified in the CTA and CP conditions. A Mann-Whitney test revealed that the CP method produced significantly more individual problems compared to the CTA method relating to layout and content problems.

Table 6.12: TA methods and individual problem type

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Navigation	4.45	1.57	4.80	2.30	U= 222.5, *z*= .618, *p*= .547
Layout*	4.00	1.86	6.10	2.90	U= 274, *z*= 2.02, *p*= .046
Content*	0.65	0.48	1.30	0.86	U= 285.5, *z*= 2.59, *p*= .020
Functionality	2.20	1.07	2.50	1.19	U= 226.5, *z*= .763, *p*= .478

* p< 0.05 significance obtained

6.7.4.2 Final Usability Problems

The CP method detected 83 final usability problems in the tested website, 10 of which were new problems that were not detected in the previous study. The CTA method, as mentioned in section 5.10.4.2, detected 60 problems on the website (see Table 6.13). Accordingly, the CP outperformed the CTA method with respect to the range of final problems detected. The percentages of unique final problems identified by CTA and CP are 13% and 37% respectively. The students applying the CTA method did not find 36 problems that were uncovered by the CP method. The students applying the CP method did not find 13 unique problems that had been uncovered by the CTA method. Note that the number of unique problems found by the CTA in the previous study on the website was 16. However, the CP method managed to find three of the 16 unique CTA problems in this study, reducing the number of unique CTA problems to 13. Both groups commonly identified 47 of the total number of problems. A list of usability problems found on the tested website is presented in Appendix E3.

Table 6.13: TA methods and the number of final problems

	# of problems	% of problems	# of unique problems	% of unique Problems
CTA	60	62 %	13	13%
CP	83	86 %	36	37%
Total	96	100 %	49	51%

Final usability problems and their sources

The final usability problems were coded according to their source—that is, the way in which they came to light: observation, verbalisation, or a combination of both (as explained in section 4.9.1.2). Table 6.14 shows the number of problems detected by the CTA and CP methods according to their problem sources. As can be seen, the CTA method detected 7 problems derived from observation evidence, 17 from verbal evidence, and 36 from a combination of the two. In the CP test, 5 problems were derived from observation evidence, 12 from verbal evidence, and 67 from a combination of the two. The CP method detected a larger number of both overlapping and unique problems from the combined sources than did the CTA method.

Table 6.14: TA methods and final problem sources

	CTA		CP	
	Unique	**Overlapping**	**Unique**	**Overlapping**
Observed	1	6	0	5
Verbalised	9	8	8	3
Both	3	33	28	39
Total	13	47	36	47

Final usability problems and severity levels

Table 6.15 sets out the number of problems according to severity level for the CTA and CP methods. The CP method managed to identify the four critical problems discovered on the site in the previous study. 31.66% (19 problems) of the final problems from the CTA method were high impact problems (with critical and major effects), and 68.33% were low impact problems (with minor and enhancement effects), whereas, for the CP condition, 18% (15 problems) of final problems were high impact. In terms of the unique problems, the results revealed that that 38% (5 problems) of the unique problems identified by the

CTA method were high impact problems. However, of the problems identified by the CP method, 9% (3 problems) were high impact problems.

Table 6.15: TA methods and final problem severity levels

	CTA		CP	
	Unique	**Overlapping**	**Unique**	**Overlapping**
Critical	0	4	0	4
Major	5	10	3	8
Minor	7	31	25	33
Enhancement	1	2	8	2
Total	13	47	36	47

Looking at the manner in which the unique problems were detected. As many as 75% (27 problems) of the problems identified by the CP method were detected through the combined source, with 91% (33 problems) of these being low impact problems. On the other hand, 23% (3 problems) of the problems detected by the CTA method were brought to light by the combined source, and all of these were major impact problems (see Table 6.16).

Table 6.16: Sources and severity levels for the unique final problems in the TA conditions

	CTA			CP		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Critical	0	0	0	0	0	0
Major	1	1	3	1	0	2
Minor	0	7	0	0	0	25
Enhancement	0	1	0	0	8	0
Total	1	9	3	1	8	27

Final usability problem types

The 10 new unique problems detected by the CP method in this study were categorised by the usability experts into 1 navigational problem, 4 layout problems, 3 content problems, and 2 functional problems. Table 6.17 sets out the number of final usability problems for each problem type for each of the TA conditions. Compared with the CTA method, the CP method identified more problems of each type, and also detected more unique problems of each type than did the CTA method.

Table 6.17: TA methods and final problem types

	CTA		CP		Total
	Unique	**Overlapping**	**Unique**	**Overlapping**	
Navigation	3	15	5	15	23
Layout	5	20	17	20	42
Content	4	5	7	5	16
Functionality	1	7	7	7	15
Total	13	47	36	47	96

Figures 6.2 and 6.3 depict the final problems detected by each TA method, displayed according to their types and severity level . As these figures show, the critical problems detected by the two methods related to navigational, layout, and functionality problems.

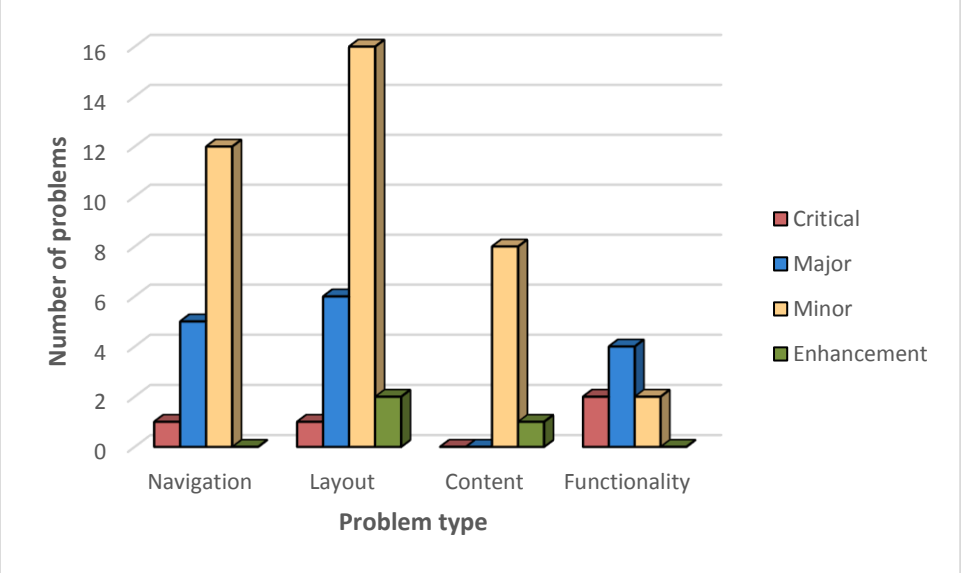


Figure 6.2: Types and severity levels for the final problems in CTA condition

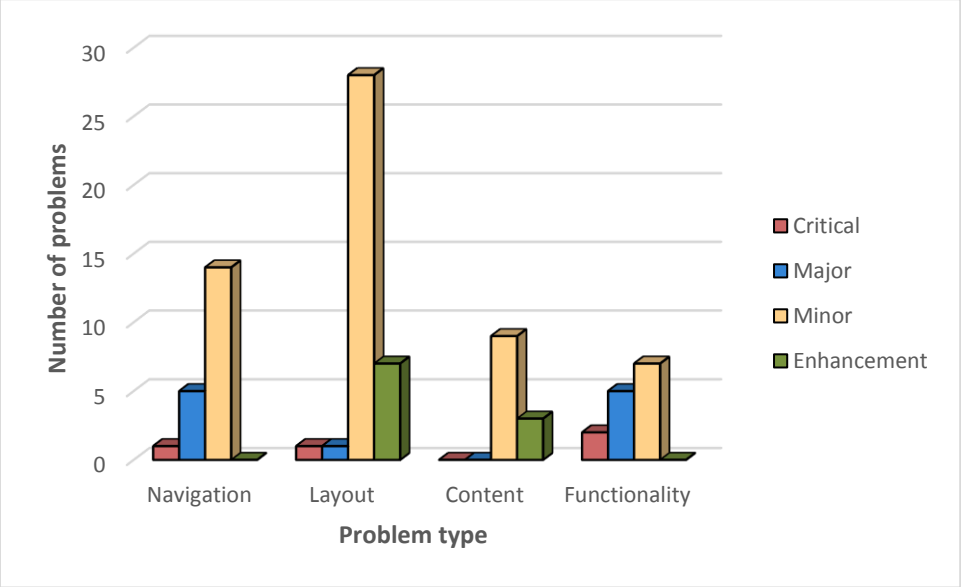


Figure 6.3: Types and severity levels for the final problems in CP condition

Table 6.18 breaks down the unique problems identified by the two methods (49 problems) according to their problem sources and types, and shows that all unique navigation, content, functionality problems identified by the CP condition were derived from combined source.

Table 6.18: Sources and types for the unique final problems in the TA conditions

	CTA			CP		
	Observed	**Verbalized**	**Both**	**Observed**	**Verbalized**	**Both**
Navigation	0	3	0	0	0	5
Layout	1	2	2	0	8	9
Content	0	4	0	0	0	7
Functionality	0	0	1	0	0	7
Total	1	9	3	0	8	28

A further examination of the types and severity levels of the unique problems is shown in table 6.19, and suggests that, for the CP condition, all problems relating to layout were at low severity levels.

Table 6.19: Types and severity levels for the unique final problems in the TA conditions

	CTA				CP			
	Critical	**Major**	**Minor**	**Enhancement**	**Critical**	**Major**	**Minor**	**Enhancement**
Navigation	0	1	2	0	0	1	4	0
Layout	0	3	1	1	0	0	9	8
Content	0	0	4	0	0	0	7	0
Functionality	0	1	0	0	0	2	5	0
Total	0	5	7	1	0	3	25	8

Reliability of problem identification and classification

As in the previous experiments, an additional evaluator was recruited to carry out an inter-coder reliability check on the usability problem analysis. The independent evaluator in this study was a PhD student under the supervision of Dr Pam Mayhew. The second evaluator independently analysed two randomly selected tests from the CP condition. The any-two agreement formula provided by Hertzum and Jacobsen (2001), explained in Section 4.9.4.2, was used to calculate the inter-coder reliability across the six videos. The average any-two agreement for the individual problem identification across the two videos was 73% (individual agreements were 73% and 72%). The any-two agreement for the final usability problems was 78%. The reliability of the coding of the problem source and severity level was examined using Cohen's Kappa (Field, 2009), explained in Section 4.9.4.2. For the individual problem levels, the resulting Kappa value for the problem source was 0.689 and for problem severity it was 0.752. For the final usability problems, the resulting Kappa value for problem source was 0.744, and the severity level was 0.832. This indicates high reliability for the coding.

6.7.5 Comparative Cost

The cost of employing the two TA methods under study was measured by recording the time expended by the evaluator on conducting tests and analysing the results for each method. As mentioned in section 3.9.4, session time refers to the time taken to carry out the entirety of each testing session (including the instruction of participants, collection of data, and solving any problems that arose during the session), and analysis time refers to the time taken to extract usability problems from each method's testing data. Session time was recorded via an observation sheet (Appendix D8), and analysis time was measured using a free web-based time tracking application called "Toggle" (Version, 2013). The

collected data from these measures was used to create a costing evaluation of the methods. The following sub-sections review the time taken for each TA method (section 6.7.5.1) and provide an estimation of their financial cost (section 6.7.5.2).

6.7.5.1 Temporal Cost

Table 6.20 shows the time spent by the researcher on applying and analysing the results for the two methods. As is clear from the table, the CP method required a longer session time (802 minutes) than the CTA method (723 minutes). The total time taken to apply the two methods was 1525 minutes.

Table 6.20: TA methods and time expense

	CTA	**CP**	**Total**
Session time *(m)*	723	802	1525
Analysis time *(m)*	865	1006	1871
Total time *(m)*	1588	1808	3396

An independent t-test was conducted to determine if there were significant differences in the mean session time between conditions. The Shapiro-Wilk test showed that the data were approximately normally distributed for the two TA groups, with $p = .087$ for the CTA group and $p = .193$ for the CP group. The assumption of homogeneity of variances was also met ($p = .529$). The test found no significant difference between the conditions with regard to session time (see Table 6.21).

Table 6.21: Session time for the TA methods

	CTA		**CP**		**Value**
	Mean	**SD**	**Mean**	**SD**	
Session time (*m*)	36.00	6.25	40.10	8.18	t=.-1.70, df= 38, $p = .096$

The total time taken to identify usability problems using the two methods was 1871 minutes, with the CP method requiring a greater amount of time (1006 minutes) in comparison to the CTA (865 minutes). An independent t-test was conducted after meeting assumptions of normality—$p = .496$ for the CTA group and $p = .461$ for the CP group—and homogeneity of variance ($p = .158$). The test showed that the analysis time for the CP condition was significantly longer than for the CTA condition. This can be explained by

the higher number of usability problems, which would lead to more time being spent on analysis and reporting (see Table 6.22).

Table 6.22: Analysis time for the TA methods

	CTA		CP		Value
	Mean	**SD**	**Mean**	**SD**	
Analysis time (m)*	43.25	4.91	50.30	7.32	t=.-3.17, df= 38, *p*= .003

* p< 0.05 significance obtained

Overall, the results showed that the CTA method required less time (1588 minutes) than the CP method (1808 minutes). The total time taken by the two methods—i.e. session time and analysis time for both CTA and CP—was 3396 minutes. By dividing the total evaluator time spent on a method by the number of problems identified by that method, the estimated temporal cost of detecting a problem can be computed and compared (Als et al., 2005). The CTA method had an estimated temporal cost of 26.46 minutes per usability problem, whereas the CP method had a cost of 21.78 minutes per usability problem (see Table 6.23).

Table 6.23: TA methods' temporal costs per problem

	Time spent (*m*)	**Problem found**	**Time per Problem (*m*)**
CTA	1588	60	26.46
CP	1808	83	21.78
All	3396	108	31.44

6.7.5.2 Financial cost

As mentioned in section 4.9.5, Martin et al. (2014) states that usability evaluators charge a rate of £800.00 per 7.5-hour day, or approximately £107 per hour. Table 6.24 shows the number of evaluator hours spent conducting tests and analysing the results for each method, multiplied by the hourly cost of a usability evaluator to produce the total financial cost of each TA evaluation (rounded to the nearest pound). It can be seen from Table 6.24 that CTA testing would cost £2831, which is less than the cost of the CP method (£3224). The cost of the application and analysis of the two methods would be £6056.

Table 6.24: TA methods' financial cost

	Evaluator Minutes	Evaluator Hours	Hourly Fee	Financial Cost
CTA	1588	26.46	£107	£2831
CP	1808	30.13	£107	£3224
All	3396	56.60	£107	£6056

By comparing the financial costs of each method against the number of problems detected, the financial cost per problem can be deduced (Martin et al., 2014) (see Table 6.25). These calculations indicate a cost of £47 per problem for the CTA method, and a cost of £38 per problem for the CP method.

Table 6.25: TA methods' financial costs per problem

	Financial Cost	Problem found	Cost per Problem
CTA	£2831	60	£47
CP	£3224	83	£38
All	£6056	108	£56

6.7.6 Relationship between Sample Size and Number of Problem Detected

As mentioned in Section 2.4.1, it has been stated—to some debate—by Nielsen (2000) that five test participants are enough to find 85% of usability issues. The first and second experiments did not achieve the results to support this claim. In this experiment, eighteen *pairs* were needed to find almost 85% of the problems. This strongly supports the argument that "five participants" can not reveal 85% of the usability problems in a given interface.

This section explores the relationship between sample size and the number of problems detected in the CP condition, comparing this with the CTA result in the previous experiment. The section first explores how the CP group performed as a whole, and then how the first five teams and the best-performing five teams did in this experiment. It finishes by determining the number of participants required to find 85% of the problems in the CP condition.

Figure 6.4 illustrates the performance of CP and CTA participants. The first 5 teams were able to discover just over 29% of the usability problems found, whereas the first five participants in the CTA condition in the previous experiment found 18% of the problems.

To find 60 usability problems, which was the total found by the CTA method in the previous experiment, the CP method needed 12 sessions, compared to 20 sessions for the CTA approach. The top-performing five couples in in the CP condition were able to detect 37% of the usability problems. By contrast, the top-performing five participants in the CTA condition found 25% of the problems. Accordingly, it can be said the CP method performed better in terms of the relationship between the sample size and the number of problems detected. Looking at the number of pairs required by the CP method in order to detect 85% of the number of the problems found. The adjusted average detection rate (explained in section 4.9.6.2) of usability problems for the CP condition was 0.094, so the sample size needed to reveal 85% of the problems is 18 pairs for the CP method compared to 33 participants for the CTA method.

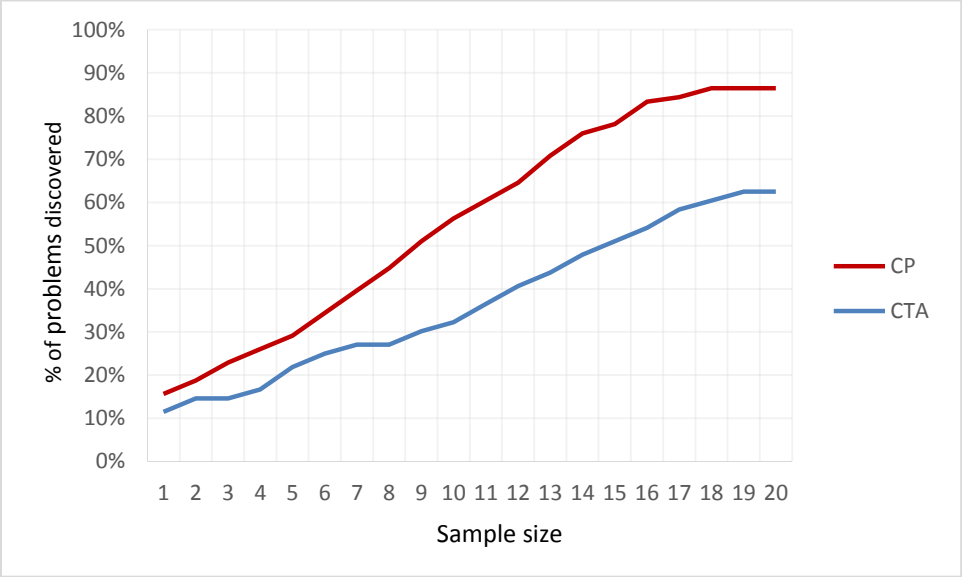


Figure 6.4: Participants' performances (cumulative) in the CP and CTA conditions

6.7.7 Correlational Analysis of Usability Measures

As in the previous two studies, the correlations between the most common usability measures—i.e. number of usability problems detected, task success rate, time spent on task, participants' satisfaction with the website (i.e., SUS score), mouse clicks, and browsed pages—were analysed using Spearman's correlation test (explained in section 4.9.7). Table 6.26 sets out the results obtained from employing the correlation test, which can be analysed as follows:

- There is a strong, statistically significant, positive relationship between time spent on tasks, mouse clicks, and visited pages in the two TA conditions.
- There is a statistically significant positive relationship between browsed pages and mouse clicks.
- There is no statistically significant relationship between the number of usability problems discovered and participant satisfaction with the website in either of the TA conditions.
- There is a statistically significant relationship between time spent on tasks and the number of usability problems discovered in the CP condition.
- There is no statistically significant relationship between task performance measures and participant satisfaction.

Table 6.26: Correlations amongst usability measures

Usability Measures		**Task success**	**Task time**	**SUS score**	**Browsed pages**	**Mouse clicks**	**Usability problems**
Task success	CTA	1	.175	.254	.128	.132	-.128
	CP	1	.261	.307	.195	.222	-.264
Task Time	CTA		1	-.239	.715**	.802**	.443
	CP		1	.086	.904**	.939**	.523*
SUS Score	CTA			1	-.178	-.258	-.354
	CP			1	-.325	-.040	-.328
Browsed pages	CTA				1	.586**	.354
	CP				1	.741**	.112
Mouse clicks	CTA					1	.246
	CP					1	.433
Usability problems	CTA						1
	CP						1

* Correlation is significant at the .005 level (2-tailed); ** Correlation is significant at the .0001 level (2-tailed)

6.8 Discussion

This section discusses the study's findings and compares them to some of the related literature. The main findings of the study are summarised in Table 6.27.

6.8.1 Think-Aloud Methods and Participants' Task Performance

The CP method did not have an impact on participants' task solving process, as the CTA and CP methods show no statistically significant differences in task solving accuracy, efficiency, or navigational behaviour. Reactivity, therefore, was not evident in the CP method. The CP participants performed their tasks neither better nor worse than the participants in the CTA condition. This corresponds to earlier findings by Adebesin et al. (2009), Als et al. (2005), and Van den Haak et al. (2004). This finding implies that practitioners have a free choice between using the traditional TA method or the CP methods if interested in measuring participant task performance.

6.8.2 Think-Aloud Methods and Participants' Experiences

The CP method seemed to elicit more positive responses from the participants than the CTA method. This finding seems to be in line with Van den Haak et al. (2004) who suggested that interaction between participants during the usability testing session could make the participants more feel comfortable and secure, therefore making them more likely to put forward their opinions. However, despite participants in the main preferring the CP method, the CP collaborators also reported that the presence of the evaluator during testing was more unnatural. This suggests it might be better for evaluators to monitor the CP test from a different room.

Regarding participants' satisfaction with the tested website, the CP method seems to have no distinguishable effect when compared to classic CTA test. This result indicates that it is legitimate to collect data regarding participants' satisfaction when using the co-participation testing.

6.8.3 Think-Aloud Methods and Usability Problems Identified

The results illustrated significant differences between classical think-aloud and co-participation on the identification of usability problems. The current experiment shows that paired participants find more usability issues than single test participants at both the individual and final problem levels. On average the pairs detected 14 usability problems over nine tasks, whereas the single participants found an average of 11 usability issues for the same number of tasks. It was also found that the CP method identified more low

severity problems relating to layout and content problems. These findings concur with Als et al. (2005) who found that paired test participants detected significantly higher number of usability problems than did single test participants. However, it contradicts Van den Haak et al. (2004) who found no such difference. This may be because in the Van den Haak study, the researchers did not consider the level of acquaintance between the pairs. In addition, the researchers did not apply a structured approach in extracting the usability problems in order to enhance the validity of data and safeguard against the evaluator effect.

6.8.4 Think-Aloud Methods and Cost

The findings of this study reveal that the CTA method costs less than the CP method in terms of the total time expended by the evaluator to conduct testing sessions and analyse results. In addition, the financial cost of the CTA method was estimated to be less than that of the CP method. This finding contradicts with Als et al. (2005) who found that the CP require less time from the evaluator than the CTA to conduct the tests and analyse the results.

6.8.5 Think-Aloud Methods and Sample Size Needed

In terms of the relationship between the sample size and the number of problems detected, the results showed that the debatable argument that five participants is enough to identify 85% of problems was not supported by this study. The results for the best performing five pairs did not exceed 37% of problems discovered. Moreover, the performance of the first five teams did not exceed 29% of the problems. The results also found that the CP method would require fewer test sessions than the CTA in order to find 85% of the problems.

Table 6.27: Overview of the main findings of the co-participation study

Results in terms of	The CP study
Task performance	
- Successful task completion	No difference between the two TA methods
- Task duration	No difference between the two TA methods
- Mouse clicks	No difference between the two TA methods
- Browed pages	No difference between the two TA methods
Participant experiences	
- The tested website	No difference between the two TA methods
- The TA method	CP method was evaluated more positively
Usability problems	
- Individual problems	
Detection means	CP produced higher number of individual problems
Source of problems	CP produced higher number of combined problems
Severity of problems	CP produced higher number of low severity problems
Types of problems	CP produced higher number of content and layout problems
- Final problems	
Detection means	CP produced higher number of final problems
Source of problems	CP produced higher number of combined problems
Severity of problems	CP produced higher number of low severity problems
Types of problems	CP produced higher number of content and layout problems
Unique problems	CTA: 13, CP: 36
Methods Cost	
- Temporal cost	CTA required less time than the CP method
- Financial cost	CTA would require less financial cost than the CP method
Sample size needed	CTA required more test sessions than the CP method to find 85% of the problems

6.9 Summary

The primary aim of this study was to assess the utility and validity of the co-participation method in comparison to the classic think-aloud method. The two methods were compared through an evaluation of a library website, which involved five points of comparison: overall task performance, test participants' experiences, number and type of usability problems discovered, the cost of employing the methods, and the relationship between the sample size and the number of problems detected.

The results of the study show significant differences between the performances of the two types of testing methods. The co-participation method was evaluated more positively by

users, led to the detection of more minor usability problems, and performed better in terms of the relationship between the sample size and the number of problems detected. The method, however, was found to require a greater investment of time and effort on the part of the evaluator in comparison to the classic method. This study found no difference between the methods in terms of task performance.

Based on the above findings, it can be concluded that the co-participation method seems to be an appropriate method for those usability practitioners who seek to find a high quantity of problems at low severity levels, or feel that it is vital that the participants in their usability test experience their participation as pleasantly as possible. Otherwise the classic method seems to be a more cost-effective method as it has the same ability of revealing high-severity problems, requires less time and effort from the evaluator, and involves rewarding one participant per test session instead of two.

The next chapter will discuss the findings of all three research studies.

7

DISCUSSION

7.1 Overview

This thesis has investigated the validity and utility of different think-aloud methods in usability testing. The findings support the assertion that think-aloud protocols are valuable evaluation tools. Regardless of the think-aloud variant used, the think-aloud methods studied and reported on in this thesis revealed a large number and wide range of usability problems.

This chapter pulls together and discusses the results of the three studies comprising this research, presented in Chapters 4, 5 and 6, by referring to the aims, objectives, and research questions of this thesis (Chapter 1), the literature review (Chapter 2), and the context of the research. Each of the research questions (section 1.5) will be addressed in turn, beginning with those relating to the validity of the think-aloud variants (section 7.2), before moving on to discuss the notion of utility (section 7.3). A set of practical implications and recommendations for evaluators looking to implement think-aloud methods in their usability testing is presented in section 7.4. Note that the results of each individual study are discussed more comprehensively at the end of each individual study chapter.

7.2 Validity

As was mentioned in section 3.9, the validity of usability testing methods refers to the degree to which the data collected conform to the real-world use of the system (Blandford et al., 2008). The validity of the think-aloud methods under investigation in this research was examined by means of the following two research questions:

Research Question 1 (RQ1): Are there discrepancies between think-aloud methods with regard to participants' task performances?

Research Question 2 (RQ2): Are there discrepancies between think-aloud methods with regard to participants' testing experiences?

7.2.1 Think-Aloud Methods and Task Performance

As mentioned in section 1.5, task performance measures were used to assess the level of reactivity for each variant. The term "reactivity" refers to a change in participants' task

performance affected by the double workload of having to perform tasks and think aloud simultaneously (Fox et al., 2011), and is often a concern when using think-aloud methods. Each of the three studies compared the think-aloud variants with respect to task performance, i.e. the number of tasks completed successfully and the time taken to complete tasks, as well as navigational behaviour (collected in Study Two and Study Three). The results of this comparison are shown in Table 7.1.

Table 7.1: Results of the three studies with respect to task performance

Study	Successful task completion	Overall task completion time	Navigational Behaviour
Classic Think-aloud	No differences	No differences	--
Relaxed Think-aloud	No differences	AI took more time than CTA & SC	AI participants visited more pages and clicked their mouse more than the CTA and SC participants
Co-participation	No differences	No differences	No differences

In Study One, three classic think-aloud methods were examined: the concurrent think-aloud (CTA), the retrospective think-aloud (RTA), and the hybrid methods (HB) in which participants think aloud both concurrently and retrospectively. The study found no differences in the participants' task performance between the classic think-aloud methods: concurrent think-aloud method, concurrent think-aloud in the hybrid method, and the silent condition in the retrospective think-aloud. Therefore, it can be stated that thinking aloud concurrently while performing tasks did not affect participants' task accuracy (section 4.9.2). This finding lends support to Ericsson and Simon's (1993) argument that thinking aloud does not have an effect on task solving. This implies that the task performance data collected when using concurrent think-aloud methods can offer an accurate representation of real-world use.

In Study Two, the performance of the classic CTA method with two relaxed variations on this method in which the evaluator plays a more active role—namely, the active intervention (AI) method and the speech-communication (SC) method—was compared. The results of the study showed that the AI method slowed down the process of task solving and led to higher numbers of mouse clicks and pages viewed when compared to the CTA and SC conditions (section 5.10.2). These findings support Ericsson and Simon's (1993) assertion that the practitioners' use of interventions could disrupt participants' thought processes, causing a change in this process and, consequently, in task performance.

Probing increases the cognitive load placed on the participant in that trying to respond to the evaluator's questioning whilst completing a task is heavily demanding in terms of attentional resources (Ericsson and Simon, 1993). This triggers alarm signals that any data collected using this type of think-aloud method might be a false representation of the user's interaction with the tested system.

In Study Three, the classic concurrent think-aloud method was compared with the co-participation method, wherein a pair of participants work together to perform their tasks, and verbalise their processes as they interact with the interface and with one another. The results of the study found no difference between the methods in terms of task performance (section 6.7.2).

In sum, the results of the comparative studies regarding task performance show signs of reactivity only in the AI method.

7.2.2 Think-Aloud Methods and Participants' Experience

In all of the three studies, questionnaires were distributed to all participants, inviting them to share their thoughts on the usability of the test objects, and their experiences of taking part in the test session.

There were four main topics in the questionnaire:
A) Satisfaction with the usability of the website used;
B) Working procedure (i.e. did the participants feel they had worked any differently from usual during the test session?);
C) Experiences with the think-aloud variant (i.e. how had the participants felt about having to think-aloud (concurrently or retrospectively) or work in teams of two?);
D) The presence of the evaluator (i.e. how had the participants felt about the presence of the evaluator?).

The results of this questionnaire for each of the studies are presented in Table 7.2.

Table 7.2: Results of the three studies with respect to participant experiences

Study	Satisfaction with website	Working procedure	Experiences with think-aloud	Presence of the evaluator
Classic Think-aloud	No differences	No differences	HB was considered more-time consuming than the CTA and RTA	No differences
Relaxed Think-aloud	No differences	AI participants felt they worked slower	No differences	AI participants felt more distracted by the evaluator than CTA and SC participants
Co-participation	No differences	No differences	CP more positive than CTA	CP collaborators felt the presence of the evaluator was unnatural

What is clear from Table 7.2 is the fact that, in each of the three studies, think-aloud methods had no effect on participants' perceptions of the usability of the chosen websites, as no significant differences were found between the methods—even though, as mentioned in section 7.2.1, significant differences in task performance were observed between the AI method and the CTA and SC methods in the second study. The explanation provided for this was the weak correlation in the studies between task performance measures and participants' satisfaction with the usability of websites, which meant that a poor task performance would not automatically result in a low level of satisfaction with the website. This phenomenon was also captured in other usability studies (Frøkjær et al., 2000; Hornbæk and Law, 2007; Nielsen and Levy, 1994b). This finding indicates that it is legitimate to collect data regarding participants' satisfaction with website usability using any of the thinking-aloud methods studied in this thesis.

With respect to the working procedure, the results of the three studies showed no significant differences between all but one of the methods, as CTA, CP, and SC participants indicated that they had not worked all that differently from usual. The AI participants, however, indicated that they had worked slower than they otherwise would have done (section 5.10.3). These experiential data seem to support the notion of reactivity associated with the AI method as a consequence of interference and extra cognitive load.

With regard to participants' experience with think-aloud methods, the results of the studies showed the participants who had worked with the CP method were more positive about the method than the participants from the other test groups: the CP participants found

working together to be more natural and pleasant (section 6.7.3). This implies that working in teams of two has a positive effect on the way in which participants perceive their participation in a usability test. This is not to say that the participants who employed other methods were negative about them, as the average scores for those variants ranged from neutral to positive. That said, the users of the HB method found the task of verbalising their thoughts both concurrently and retrospectively to be time-consuming (section 4.9.3).

Regarding the presence of the evaluator, the results in Table 7.2 show that this seems to have had a more detrimental effect in the AI condition, with participants indicating that the presence of the evaluator was distracting. The negative effect of the evaluator's presence on participants in the AI condition was attributed to the evaluator's probing (section 5.10.3). Interestingly enough, in the third study, CP collaborators found the presence of the evaluator to be more unnatural than did the CTA participants. This may suggest that when employing the CP method, it might be more appropriate if evaluators monitor the test from a different room or remotely.

Overall, the results for the second research question indicate that participants in general preferred the CP method, and that the ecological validity (i.e. the extent to which to a method is comfortable for participants to use) was compromised in the AI method. In other methods no strong preferences or risks to validity were observed. This suggests that, if Ericsson & Simon's (1993) guidelines for minimum interaction between experimenter and participants are observed, the discomfort reported by the AI participants can be avoided.

7.3 Utility

As was mentioned in section 3.9, the utility of usability evaluation methods refers to the usefulness of a method in assisting usability work (Blandford et al., 2008). In this thesis, the utility of the think-aloud variations was investigated using the following three research questions.

Research Question 3 (RQ3): Are there discrepancies between think-aloud methods with regard to the quantity and quality of usability problems they detect?

Research Question 4 (RQ4): Are there discrepancies between think-aloud methods with regard to the cost of employing the methods?

Research Question 5 (RQ5): Are there discrepancies between think-aloud methods with regard to the relationship between sample size and number of problems detected?

7.3.1 Think-Aloud Methods and Usability Problems

Each of the three studies compared the think-aloud variants in terms of number and quality of problems detected. The number of problems, considered in terms of the manner of their detection (i.e. by means of observation, verbalization or a combination of both) as well as their severity, will be discussed first, followed by the types of problems detected and uniqueness. Table 7.3 shows the results of all three studies regarding the quantity and quality of problems detected.

Table 7.3: Results of the three studies with respect to usability problems

Study	Number of problems	Source of problems	Severity of problem	Types of Problem	Uniqueness of problems
Classic Think-aloud	CTA and HB revealed more problems than RTA	CTA and HB revealed more verbalized problems	CTA and HB revealed more minor problems	CTA and HB revealed more layout problems	CTA and HB revealed more unique problems
Relaxed Think-aloud	No differences	No differences	AI revealed more enhancement problems than CTA	CTA revealed more content problems than SC	CTA and AI revealed more unique problems
Co-participation	CP revealed more problems than CTA	CP revealed more combined problems	CP revealed more minor & enhancement problems	CP revealed more layout and content problems	CP revealed more unique problems

As is clear from Table 7.3, the CTA and HB methods were more productive than the RTA method in the first study in terms of the quantity of usability problems found (see section 4.9.4). This finding supports Ericsson and Simon's (1993) argument that potential information may be lost when employing the RTA method. However, it does not lend support to Ericsson and Simon's (1993) claim that collecting both concurrent and retrospective data can positively affect the richness of data collected. This, as mentioned in section 4.9.4, might be attributable to the HB participants feeling that they had already

provided detailed comments in the concurrent phase and not wishing to repeat themselves. Participants may also have been feeling tired due to the extended session time.

In Study Two, no difference between the three think-aloud variants was found. Thus, in terms of quantity of output, the think-aloud variants can be said to be similar (section 5.10.4). Nevertheless, when the CTA method was compared with the CP method in the third study, the latter method proved more fruitful with respect to the number of problems detected (see section 6.7.4). This result was explained by the fact that the teams in the CP condition had two pairs of eyes which might allow them to uncover more problems on the interface.

The second part of the research question concerns the quality of problems detected by the think-aloud variants.

With respect to the manner in which problems were detected, in the classic think-aloud study, the CTA and HB methods revealed more verbalized problems than the RTA method (section 4.9.4). A possible explanation for this difference is that asking test participants to report problems after performing tasks silently (rather than concurrently, whilst working) may increase the likelihood of their simply forgetting to report problems during the retrospective phase, even if they had noticed these problems whilst working. A second difference between the think-aloud variants investigated in this thesis concerned the CP method. In the third study, the CP participants detected more problems from a combination of observation and verbalization than the CTA participants (section 6.7.4); a finding which, like the above-mentioned difference in the number of problems, could be explained by the fact that there were two people involved in the CP condition, and so there are likely to be more opportunities for problems to be detected. In Study Two, no difference between the three think-aloud variants in terms of the manner of problem detection was found.

In terms of the severity of problems detected, in the first study, the CTA and the HB methods found a greater number of minor problems than did the RTA method (section 4.9.4). In the second study, the AI method produced more enhancement problems than the CTA method—however, this represents a small proportion of the problems detected in the studies (section 5.10.4). In the third study, the CP method detected more minor and enhancement problems than the CTA method (section 6.7.4).

Addressing now the types of problems that were detected. The three experimental studies undertaken during this research show that navigation and layout presented the most problems to the users of the tested library websites. Research evaluating the usability of online libraries has yielded similar results, indicating that navigational and layout problems are among the most frequently encountered problems by users (e.g. Imler and Eichelberger, 2014; Bull, Craft, and Dodds, 2014). In the first study, the CTA and HB methods produced significantly higher numbers of layout problems than the RTA method. The CTA method was determined in Studies One and Two to be the most cost-effective method; however, in the third study, the CP method outperformed the CTA, revealing a higher number of layout problems. There were also some differences between the CTA, SC, and CP methods regarding content problems. The CTA method uncovered a higher number of content problems than did the SC in the second study, but the third study found that the CP method was once again more successful than the CTA in spotting problems relating to content. However, these differences primarily concerned very small numbers of problems. As such, it could be argued that the CTA, SC, and CP variants are comparable with respect to their ability to detect content problems.

The problems detected by the think-aloud variants in each study were also analysed with respect to uniqueness. Results, as shown in Table 7.3, indicate that the RTA and SC methods were the least effective at detecting unique problems, while the CP detected higher quantity of unique problems than the CTA.

Overall, the results for the third research question suggest that the CP method is the most profitable among the think-aloud variants with respect to the number of problems detected. The method also identified more problems from the combined source. However, most of the problems found were low severity problems. Given that one of the main tasks of usability practitioners is to prioritise problems for developers to address, it is therefore reasonable to suggest that the CP method is appropriate for those usability practitioners who seek to find a high quantity of low-severity problems. Otherwise the CTA method, which shows similar capabilities for detecting high-severity problems, should be utilised. The utility of relaxed think-aloud methods in improving the usability problem sets is not supported by these studies.

7.3.2 Think-Aloud Methods and Cost

Table 7.4 presents the results of the three studies with respect to cost. As can be seen, the CTA method required the shortest time among the think-aloud methods. It was also estimated to incur a lower financial cost than the other variations. On the other hand, the HB method required the longest time from the evaluator, and would thus be the most expensive method to apply. The SC method was ranked second, following the CTA method, in terms of both the temporal and financial costs. This method was followed by the CP method, which required shorter testing and analysis times from the evaluator, and was estimated to incur a lower financial cost than both the AI and the RTA methods. However, it should be bear in mind that the CP involves rewarding two participants per test session which means the CP method is more expensive than the single-participant methods in this respect.

Table 7.4: Results of the three studies with respect to cost

Study		Temporal cost (*m*)	Financial cost (£)
Classic Think-aloud	CTA	1373	£2448
	RTA	2245	£4002
	HB	2383	£4248
Relaxed Think-aloud	CTA	1588	£2831
	SC	1687	£3007
	AI	1824	£3252
Co-participation	CP	1808	£3224

7.3.3 Think-Aloud Methods and the Relationship between Sample Size and Number of Problems Detected

As was mentioned in section 1.5, the last research question in this study explores the relationship between sample size and the number of problems detected, and in particular seeks to investigate whether sample sizes work differently for the TA methods under investigation. The first study found the CTA and HB method to show similar patterns. Both outperformed the retrospective method in this regard: the RTA method required considerably more test participants than the CTA and HB methods in order to find an equal percentage of problems (section 4.9.6). The second study showed that there were no differences between the three usability test variants regarding the relationship between sample size and number of problems detected: the CTA, SC, and AI conditions all behaved similarly (5.10.6). In the last study, the results showed that the CTA method would require

more test sessions than the CP method in order to detect a similar percentage of problems (6.7.6) (see Table 7.5).

Table 7.5: Results of the three studies with respect to relationship between sample size and problems

Study	Overall relationship	Sample size required to find 85% of the final number of problems
Classic Think-aloud	CTA and HB performed better than RTA	RTA required considerably more test participants (46) than the CTA (34) and HB (30) methods
Relaxed Think-aloud	No differences	No differences
Co-participation	CP performed better than CTA	CP method required fewer test participants (18) than the CTA (33)

A number of studies have claimed that five participants ought to be enough to reveal 85% of usability problems (e.g. Nielsen, 2000; Virzi, 1992). There is, however, research that argues the opposite viewpoint (e.g. Molich et al. 2004; Bevan et al., 2003) (see Table 7.6). This research engaged thoroughly with this controversial argument, as discussed in sections 4.10.5, 5.11.5 and 6.8.5. The results of the first five and best-performing five participants in each test group were analysed in order to highlight any similarities or differences between the performances of the methods. It was found that no group of five was able to detect more than 43% of the reported usability problems. In order to detect 85% of the problems, the RTA would require the highest number of test participants (46 participants). In contrast, the CP method would require 18 sessions to find the same percentage of usability issues (see Table 7.5). Table 7.6 compares the results obtained from this research with a number of previously published results. This research's findings suggest that in order to achieve satisfactory results, five participants are not enough. It may be that the complexity of websites such as online libraries is much greater than the complexity of the systems used to derive Nielsen's (2000) model, and that it is helpful to use (considerably) larger samples than those suggested by Nielsen (2000). In addition, library websites target a wide range of users, with varying user behaviours and characteristics, so more usability problems are expected to appear.

Table 7.6: Comparisons of five participants' performances in different studies

Study	Percentage of problems	Comments
Nielsen (2000)	85%	Five CTA participants
Virzi (1992)	80%	Five CTA participants
Bevan et al. (2003)	35%	Five CTA participants
Faulkner (2003)	55%	Five CTA participants
Molich et al. (2004)	75%	The top team was able to reveal this percentage
This research	17% - 43%	The range across the best and first performing five participants' results

As might be expected, the above discussion of the validity and utility of the think-aloud variants in question leads to some practical implications for usability evaluators and researchers to take into account. These implications as well as various recommendations related to the use of the think-aloud methods will be discussed in section 7.4.

7.4 Practical Implications and Recommendations

Having discussed the degree of validity and utility of the think-aloud methods in sections 7.2 and 7.3, the present section will offer various practical implications and recommendations regarding the think-aloud methods investigated in this thesis, and their utility for the evaluation of websites (see Table 7.7 and Figure 7.1).

- The varying effects of the different think-aloud methods should be considered seriously, as the findings suggest that results may differ depending on the method used. Therefore, practitioners should consider the pros and cons of think-aloud methods when deciding on a think-aloud method.
- When documenting think-aloud protocol, it is recommended that, rather than writing a vague statement such as "we had participants think aloud", practitioners describe the methods used and procedures followed in detail.
- This research highlights that practitioners have a free choice between using the traditional CTA, the RTA, the SC, or the CP methods if they wish to capture user performance in the "real context of use", as these methods do not show any effect on task performance. However, the AI method has negative effects on user performance. This triggers alarm signals that data collected using this method might be a false representation of the user's interaction with the tested system.

- Ericsson and Simon's guidelines for interaction should be followed in collecting think-aloud data. There should be minimal interaction between evaluator and participants to avoid effecting participants' task performance.
- Be aware of the possible negative effect of the AI method on participants' testing experience.
- Consider using CP when it is vital that the participants in their usability test experience their participation as being as pleasant and natural as possible.
- For CP tests, the evaluator should be located in a separate monitoring room in order to ensure the ecological validity of the test. Based on the questionnaire data, it was obvious that the CP helpers found the presence of the evaluator unnatural.
- Practitioners who are interested in detecting as many problems as possible, regardless of the quality of these problems, may wish to opt for the CP variant.
- Consider using the CP method when interested in finding higher numbers of low severity usability problems—particularly those relating to layout.
- Consider using the CTA method when seeking to identify high severity usability problems, as this research suggests that the CTA method detects similar numbers of high impact problems to the CP method, whilst incurring a lesser temporal and financial cost.
- The research shows that practitioners can collect data on participants' satisfaction with test objects using any of the think-aloud methods studied in this thesis, as there were no statistically significant differences between the conditions.
- Usability practitioners should be aware of the fact that participants' satisfaction with the perceived usability of test objects does not correlate with actual usability measures. This implies that user satisfaction should not be used as a sole metric for determining the usability of the tested interface.
- Another practical aspect that usability testers should take into account when planning to conduct RTA, HB, AI or CP tests is that the methods require a longer time for the application and analysis of the results than the classic CTA method. These methods are also estimated to cost more than the CTA method.
- This research finding's support the growing body of thought that argues the "magic number" of five participants is not sufficient to reveal an adequate number of usability problems. Therefore, practitioners, who are interested in detecting as

many problems as possible using think-aloud methods, should consider recruiting a much higher number of test participants.

Table 7.7: Research recommendations

If usability practitioners/researchers are interested in	Use	Avoid
Capturing user performance in the "real context of use"	CTA, RTA, SC, CP	AI
Capturing user performance in the "real context of use" with limited time and budget	CTA	Other methods
Finding usability problems	CTA, CP, SC	RTA, AI
Finding as many usability problems as possible, regardless of the cost of methods	CP	Other methods
Finding as many usability problems as possible, regardless of the quality of these problems	CP	Other methods
Finding as many usability problems as possible with less number of test sessions	CP	Other methods
Finding as many usability problems as possible with limited time and budget	CTA	Other methods
Finding high severity usability problems with limited budget and time	CTA	Other methods
Finding as many usability problems as possible with less number of test participants	CTA	Other methods
Measuring user satisfaction	Any method	No method
Measuring users satisfaction with limited time and budget	CTA	Other methods
Measuring users satisfaction and ensuring that the participants in their usability test experience their participation being as pleasant and natural as possible	CP	Other methods

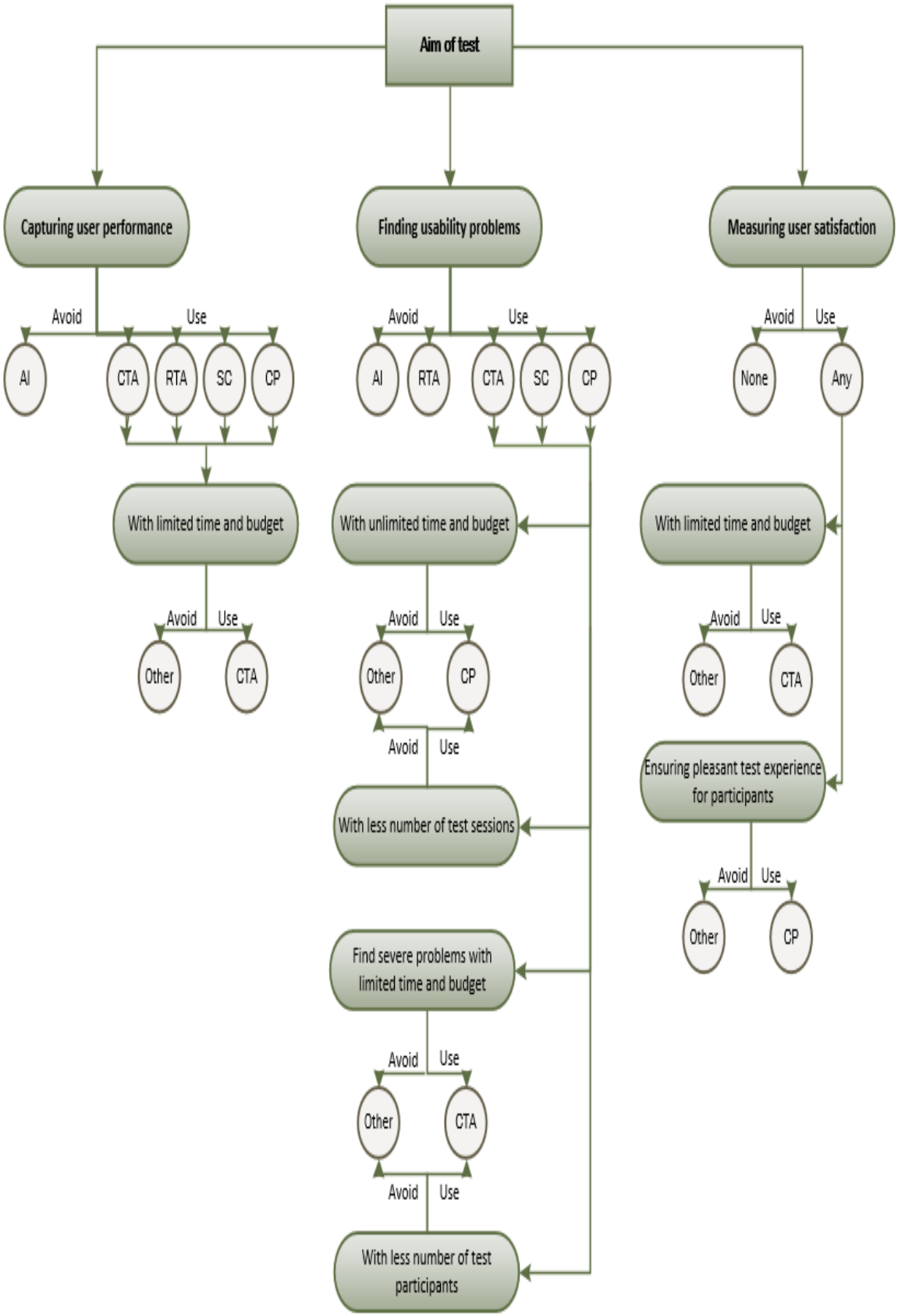


Figure 7.1: Research recommendations

7.5 Summary

This chapter has pulled together and discussed the results of the three experimental studies comprising this research, i.e. the classic think-aloud study, the relaxed think-aloud study, and the co-participation study, by referring to the aims, objectives, and research questions of this thesis (Chapter One). It has also justified the findings, linking them with previous work. A set of practical implications and recommendations for usability regarding think-aloud methods was presented in section 7.4. The next chapter intends to draw conclusions from this research.

8

RESEARCH 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 Research Summary

Usability is increasingly recognised as an important factor in the design and development of websites and web interfaces. Several studies have reported the benefits of a strong commitment to usability throughout the development life cycle of a product. Amongst the observable benefits of more usable user interfaces, are increases in performance, safety, security, user productivity, and user satisfaction. Evaluation methods which can be used to effectively assess and improve the usability of a product are therefore of critical importance. One of the most widely used methods of evaluating the usability of websites is the thinking aloud protocol, wherein users are encouraged to verbalise their experiences, thoughts, actions, and feelings whilst interacting with the design. This is intended to give evaluators direct insight into the cognitive processes employed by users as they work with an interface. These insights can then be measured, and analysed, and the data used to improve the product's usability. Despite the common usage of TA protocol in the field, the specific think-aloud procedures employed can vary widely.

This research aimed to investigate the validity and utility of the different variations of think-aloud usability testing methods. To this end, three empirical studies were conducted, using library websites, to compare the practical benefits of the various methods. The studies measured five points of comparison: overall task performance, the experiences of the test participants, the quantity and quality of usability problems discovered, the costs of employing the method in question, and the relationship between sample size and the number of problems detected. Given the research's focus on investigating different variants of the think-aloud method, and the fact that think-aloud methods are typically applied in usability laboratory settings (Norman and Panizzi, 2006), an experimental approach was used in this research.

Study One (classic think-aloud study) examined three classic think-aloud methods: concurrent think-aloud (CTA), retrospective think-aloud, and a hybrid method. In accordance with Ericsson and Simon's (1993) guidelines, the role of the evaluator was strictly non-interactive: the evaluator only intervened to remind participants to think aloud if they stopped verbalising their thoughts during testing. 60 participants were recruited for this study, with 20 participants allocated to each testing method. The numbers of participants, numbers of tasks, laboratory used, test object, and evaluation criteria were the same in each group. Only the TA methods varied between groups, as this is the issue under study. The results revealed that the concurrent method outperformed both the retrospective method and the hybrid method in facilitating successful usability testing. It detected higher numbers of usability problems than the retrospective method, and produced output comparable to that of the hybrid method. The method received average to positive ratings from its users, and no reactivity was observed. In addition, this method required much less time and effort on the evaluator's part than did the other two methods, which involved double the testing and analysis time. Lastly, in terms of the relationship between the sample size and the number of problems discovered, the concurrent and the hybrid methods showed similar patterns, and both outperformed the retrospective method in this regard. These findings suggest that the concurrent method is more effective than the retrospective and hybrid methods. A more detailed discussion of this study can be found in Chapter 4.

Study Two (relaxed think-aloud study) compared the performance of the classic CTA method with two relaxed variations on this method in which the evaluator plays a more active role—namely, the active intervention (AI) method and the speech-communication (SC) method. The second study therefore involved three groups, each consisting of 20 participants. As with the first study, all conditions were identical; only the TA method employed varied between groups. The results showed that these three methods enabled the diagnosis of a similar number of usability problems and types, and showed similar patterns with regard to the relationship between the sample size and the number of problems discovered. However, the active intervention method was found to cause some reactivity, modifying participants' interactions with the interface, and negatively affecting their feelings towards the evaluator. The AI method also required much greater investment than did the other two methods, both in terms temporal and financial cost. In this study, the SC method provided broadly similar results to those yielded by the CTA method; however previous research into the SC method has shown that the evaluator's tone of voice, attitude,

friendliness, and confidence may affect participants' subsequent verbalisations (Boren and Ramey, 2000). The results of the study therefore indicate that the supposed benefits of relaxed think-aloud methods do not seem to outweigh the risks. This study is discussed more thoroughly in Chapter 5.

Study Three (co-participation study) compared the classic CTA method with the co-participation (CP) method, wherein a pair of participants work together to perform their tasks, and verbalise their processes as they interact with the interface and with one another. This study involved a group of 40 participants working in pairs. As in the first and second studies, conditions were identical for both groups except for the TA methods used. This study found no difference between the methods in terms of task performance. However, the co-participation method was evaluated more positively by users in comparison with the classic method. It led to the detection of more minor usability problems, and performed better in terms of the relationship between the sample size and the number of problems detected. The co-participation method was, however, found to require a greater investment of time and effort on the part of the evaluator. As a result, practitioners who are interested in detecting as many minor usability problems as possible are advised to opt for the co-participation variant. Otherwise they should consider using CTA, as it has a similar efficacy in detecting high impact usability problems, and costs less. A more detailed discussion of this study can be found in Chapter 6.

8.3 Evaluation of Research Aim and Objectives

As mentioned in section 1.4, in order to accomplish the aim of the research, six objectives must be achieved. The following is an evaluation of the degree to which these were achieved.

The first objective of this research was to *explore the current and relevant literature on usability testing and think aloud protocols*. This was accomplished by reviewing the studies that have contributed to this research field, as shown in Chapter Two. The second objective of this research was to *plan a series of empirical studies which endeavour to meet the aim of the project*. This was fully accomplished as described in the Research Methodology Chapter. The third objective of this research—to *carry out the planned studies*—was achieved through conducting the three experiments outlined in the previous

section. The fourth objective of this research was to *analyse and compare of the results of the think-aloud methods investigated*, and was accomplished using figures and tables to make the comparison of the methods easy and clear, as seen in Chapters Four, Five, and Six. The fifth objective, *to discuss the findings and draw conclusions in terms of the research questions*, was met in Chapter Seven. The final objective, *to provide a set of recommendations for the benefits of future researchers, as well as for usability practitioners and engineers considering TA methods for evaluating the usability of websites*, is accomplished in section 7.4. Accordingly, it can be said that the six objectives proposed to achieve the research aim have been achieved.

8.4 Research Contributions

From the research process, several methodological and theoretical contributions have emerged, which offer a range of potential benefits. What follows is a discussion of these contributions.

The research findings contribute to the general field of website usability evaluation. They provide academics and practitioners with information on the validity and utility of the most commonly used think-aloud usability testing methods in the field. While research into think-aloud approaches has been ongoing for a number of years, the work presented here is the first to carry out a holistic comparative examination of the different variations available to professionals. The research also defines explicit operational criteria and strategies to measure the validity and utility of the investigated methods, as discussed in section 3.9. In addition, previous research has been criticised for a narrow focus on the number of problems identified by a method, which is an overly reductive means of measuring a method's utility (Wixon, 2003; Hornbeek, 2010). This research therefore employed a more robust set of assessment criteria, which included investigating the source, severity, types, and uniqueness of individual and final problems. Furthermore, the thesis provides a thorough explanation of the usability problem extraction approach (see section 3.10), which was of direct practical use in enhancing the reliability of the resultant data. This explication can be directly applied and improved by other researchers—a big step forward, given that some researchers argue that the majority of think-aloud research does not document its problem extraction methodology at all (Howarth et aI., 2009; Hornbeek, 2010).

The research has also made a theoretical contribution by testing the applicability of distinctive think-aloud models within the context of extant usability testing frameworks, such as those designed by Ericsson and Simon (1993), and Boren and Ramey (2000). Furthermore, web developers aiming to create and maintain successful websites—particularly university library websites—can also benefit greatly from the findings of this research, which explicates some of the usability issues commonly faced by users of such websites. By contributing to the improvement of the design and quality of a website, this research will also promote a better relationship between the users of a website and its administrators or owners.

Last but not least, a rather personal outcome of conducting this research is the progress made by the author towards becoming a usability professional. Nielsen (2002) stated that "to reach the goal of making technology truly suited for humans; the world will need about half a million new usability professionals over the next 20 years. The sooner their training begins, the better off we'll all be". With the experience gained from conducting this research, the author of this thesis is one step closer to becoming a usability professional. The research process has enabled the author to develop his skills and knowledge through planning and conducting a series of usability evaluation studies. This involved recruiting participants; selecting and designing tasks; selecting appropriate usability measures; conducting the usability tests; and analysing and reporting the results. The author has also published 4 papers and 2 posters in the course of this research, and participated in the annual Postgraduate Research Day in the School of Computing Sciences at UEA (see Appendix F). This demonstrates the willingness and ability of the researcher to communicate and share knowledge with other professionals in the field. Attending conferences was very fruitful in terms of both getting feedback on current research and observing the research being carried out by others in the field. It was also useful for building a strong network of links with other researchers, practitioners, and institutes.

8.5 Research Limitations

As with any project of this sort, this thesis has a number of inevitable limitations that could be improved in future work. First, the usability test sessions were performed in a formal laboratory-based setting, an important aspect for observation and analysis of results in a scientific setting. However, this sort of setting is not reflective of the environments in

which people typically access the web, and therefore might not have completely captured the normal web browsing behaviour of the participants. The second limitation concerns the demographic characteristics of the participants. While the researcher did ensure, in all evaluations, that the participants were evenly divided over the methods with respect to their demographic characteristics, they were nevertheless all drawn from one specific target group, i.e. University students. While this factor has not hindered the present research, as students represent the main target group of the test objects, it may serve to limit the application of the results to other groups who also make use of the test object, such as faculty and employees. Third, all the participants in the study were also from the same young age group, of a similar educational background, and possessed a similar level of familiarity with the Internet. This might also minimise the utility of applying the results to a broader range of users (e.g., users with low Internet experience or without an academic background, older web users, or children of school age). Fourth, the think-aloud methods in this research were only applied to university library websites. Testing different websites with different kinds of users, such as websites aimed at elderly people, may yield results that are different from the ones presented in this thesis. It seems possible, for instance, that thinking aloud while performing tasks might present greater difficulties for elderly people than for students who have grown up with web technologies. As such, testing websites with various target groups would be very worthwhile. The final limitation is the potential bias that may have been introduced by the author in conducting the usability testing. Clemmensen et al. (2009) suggest that the cultural background of the evaluator is likely to have an impact on usability testing results. Since the author is of a different nationality to the participants, there is the possibility that participants' behaviour and think-aloud data might have been influenced by cultural differences and barriers. However, the author has lived in the UK for many years—a factor which might mitigate against this limitation.

8.6 Directions for Future Research

This research has been useful, but is certainly not the "final word" on usability testing methods. There is scope for further research, most notably regarding those areas that fell outside the scope of the studies or that could have been addressed in a different way. These areas are discussed below.

As discussed in section 8.3, all participants in the three comparative studies came from a similar population—university students—and they all tested the same type of website—i.e., online libraries. With this in mind, it would be useful to replicate the studies with different types of participants or different testing interfaces to see if effectiveness of a method can vary according to these factors. It would also be interesting to replicate the studies in participants' own environment to determine if such factor can impact on the results obtained. Another suggestion for future research concerns the co-participation method. It would be of interest to compare different team compositions, such as teams of participants who are acquainted with each other versus teams of participants who have never met before, or mixed gender teams versus all-male or all-female teams. Additionally, as we have seen, the results of the co-participation study show that the participants found the presence of the evaluator unnatural. It would be interesting to experiment with the role of the evaluator during co-participation testing—for example, by comparing the results of a test in which the evaluator remains in the test room with another in which the evaluator monitors the test from a separate room. There is also scope for looking at ways to improve the retrospective think-aloud method. In the first study of this research, the retrospective think-aloud participants were presented with a video recording of their performance, on the basis of which they were asked to verbalise their thoughts retrospectively. The result of this approach showed that much potentially useful information was lost in the retrospective verbalisations. A recent trend in retrospective think-aloud testing is the placement of eye-tracking equipment (Elbabour, 2015). However, few, comparisons have been made between RTA verbalisations produced on the basis of eye tracking and other think-aloud methods. There is much potential for research in this area.

As is clear from the above suggestions, think-aloud protocols form an interesting and fruitful area for research. There are various practical and theoretical issues regarding this usability method that have not yet been investigated or that deserve more methodological investigation.

8.7 Summary

In summary, this research has provided a more holistic view than that currently available in the literature on the validity and utility of think-aloud usability testing methods. This was achieved by taking a broader, comparative focus, considering various issues and

measures. It is clear from the results presented in this thesis that Ericsson and Simon's (1993) classic concurrent think-aloud method should be employed when collecting usability data from users of online libraries, not only because it outperformed the retrospective and the hybrid methods in the first study, nor because it was shown to be more effective and valid than the relaxed methods in the second study, but because it has a similar efficacy as the co-participation method in detecting high-severity usability problems, whilst being more cost-efficient than that method. However, the co-participation method should be adopted if usability practitioners are attempting to find as many usability problems as possible, regardless of the type or severity of the problems and the cost of the test.

REFERENCES

Aaron, M. (2005). User interface design's return on investment: Examples and statistics. In Randolph G. Bias, Deborah J. Mayhew (Eds.), *Cost–Justifying Usability: An Update for the Internet Age.* Burlington: Morgan Kaufmann Press. pp. 17–39.

Adebesin, T.F., De Villiers, M.R. and Ssemugabi, S. (2009). Usability testing of e–learning: an approach incorporating co–discovery and think–aloud. In John McNeill, Shaun Bangay (Eds.), SACLA '09: *Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association*. New York: ACM. pp. 6–15.

Albert, W. and Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Oxford: Newnes.

Alhadreti O., O., Mayhew, P. J., Alshamri, M (2011). A comparison of in–lab and synchronous remote usability testing methods: effectiveness perspective. In Katherine Blashki (Ed.), *Proceedings of the IADIS International Conference on Interface and Human Computer Interaction* (IHCI*).* Rome, Italy 20–26 July 2011. Lisbon: IADIS. pp.??–??

Als, B.S., Jensen, J.J. and Skov, M.B. (2005). Comparison of think–aloud and constructive interaction in usability testing with children. In Mike Eisenberg, Ann Eisenberg (Eds.), *Proceedings of the 2005 Conference* on *Interaction Design and Children. IDC 2005*. Boulder, Colorado. 8–10 June. New York: ACM. pp. 9–16.

Alshamari, M. and Mayhew, P. (2008). 'Task design: its impact on usability testing*'*. In *The Third International Conference on Internet and Web Applications and Services*, Athens, 8–13 June. New York: IEEE. pp. 583–589.

Amaratunga, D., Baldry, D., Sarshar, M. and Newton, R. (2002). Quantitative and qualitative research in the built environment: application of "mixed" research approach. *Work Study*, 51(1), 17–31.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. and Stage, J. (2007). What happened to remote usability testing?: an empirical study of three methods. In Mary Beth Rosson,

Arms, Y. (2000). *Digital libraries*. Cambridge: MIT Press

Archer, N.P. and Yuan, Y. (1995). Comparing telephone–computer interface designs: Are software simulations as good as hardware prototypes? *International Journal of Human–Computer Studies*, 42(2), 169–18

Barendregt, W., Bekker, M. M., Bouwhuis, D. G. and Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and after some practice. *Intemational Joumal of Human–Computer Studies.* 64(9). 830–846.

Barkhuus, L. and Rode, J.A. (2007). From mice to men—24 years of evaluation in CHI. In Mary Beth Rosson, David Gilmore (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM. pp. 1–16.

Barnum, C.M. and Dragga, S. (2001). *Usability testing and research*. Boston: Allyn and Bacon, Inc.

Barnum, C.M. (2002). *Usability Testing and Research.* New York: Longman.

Barnum, C.M. (2011) *Usability testing essentials: Ready, set… test!* Amsterdam: Elsevier/Morgan Kaufmann.

Bevan, N. and Macleod, M. (1994). Usability measurement in context. *Behaviour and Information Technology*, 13(1–2), 132–145.

Bevan, N., Barnum, C., Cockton, G., Nielsen, J., Spool, J. and Wixon, D. (2003). The "magic number 5": is it enough for web testing? In Gilbert Cockton, Panu Korhonen (Eds.), *CHI '03: Extended Abstracts on Human Factors in Computing Systems,* Fort Lauderdale, Florida, New York: ACM. pp. 698–699.

Blandford, A., Hyde, J., Green, T. and Connell, I. (2008). Scoping Analytical Usability Evaluation Methods: A Case Study. *Human–Computer Interaction.* 23(3), 278–327.

Boothe, C., Strawderman, L. and Hosea, E. (2013). The effects of prototype medium on usability testing. *Applied ergonomics*, 44(6), 1033–1038.

Bošnjak, S. (2001). The Declaration of Helsinki – the cornerstone of research ethics. *Archive of Oncology,* 9(3), 179–184.

Boren, T. and Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261–278.

Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research*. London: Kogan Page.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, B. Weerdmeester (Eds.), *Usability Evaluation in Industry*. Abingdon: Taylor and Francis. pp. 189–194.

Bryman, A. (1998*).* Quantitative and qualitative research strategies in knowing the social world. In Tim May, Malcom Williams (Eds.), *Knowing The Social World.* Maidenhead: Open University Press. pp. 138–156

Bull, S., Craft, E., and Dodds, A. (2014). Evaluation of a resource discovery service: *New Review of Academic Librarianship*, *20*(2), 137–166.

Bygstad, B., Ghinea, G. and Brevik, E. (2008). Software development methods and usability: Perspectives from a survey in the software industry in Norway. *Interacting with computers*, 20(3),375–385.

Catani, M.B. and Biers, D.W. (1998). Usability evaluation and prototype fidelity: Users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42(19). New York: SAGE. pp. 1331–1335

Cairns, P. (2007).. HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it–Volume 1*. Swindon: British Computer Society. pp. 195–201

Capra, M.G. (2006). *Usability problem description and the evaluator effect in usability testing.* Unpublished: Virginia Polytechnic Institute and State University. PhD.

Chilana, P.K., Wobbrock, J.O. and Ko, A.J. (2010). Understanding usability practices in complex domains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM. pp. 2337–2346.

Choudrie, J., Ghinea, G. and Songonuga, V.N. (2013). Silver surfers, e–government and the digital divide: An exploratory study of UK local authority websites and older citizens. *Interacting with Computers*.

Clemmensen, T. and Leisner, P. (2002). Community knowledge in an emerging online professional community: the interest in theory among Danish usability professionals. *tc*, 45(3815), 2389.

Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q. and Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers*, 21(3), pp. 212–220.

Cohen, L., Manion, L. and Morrison, K (2007). *Research Methods in Education*. New York: Routledge .

Cockton, G and Lavery, D. (1999). A framework for usability problem extraction. In *Proceedings of the* 1999 *IFIP TC13 International Conferences on Human– Computer Interaction.* INTERACT'99. Amsterdam: IDS Press, pp. 344–352.

Cooke, L. (2010). Assessing concurrent think–aloud protocol as a usability test method: a technical communication approach. *IEEE Transactions on Professional Communication.* 53 (3), 202–215.

Cotton, D. and Gresty, K. (2006). Reflecting on the think–aloud method for evaluating e–Iearning. *British Journal of Educational Technology.* 37 (1), 45–54.

Creswell, J.W. (2009) *Research Design*. New York: Sage.

David Gilmore (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM. pp. 1405–1414.

Dawson, Catherine. (2002). *Practical research methods: a user–friendly guide to mastering research techniques and projects*. Begbroke: How To Books Ltd.

De Jong, M.D. and Lentz, L.R. (1996). Expert judgments versus reader feedback: A comparison of text evaluation techniques. *Journal of technical writing and communication*, 26(4), 507–519.

Denscombe, M. (2007). *The Good Research Guide for Small–Scale Research Projects, 3rd edition*. Buckingham: Open University Press.

Dewberry, C. (2004). *Statistical methods for organizational research: Theory and practice*. Hove: Psychology Press.

Dillon, A. (2001). The evaluation of software usability. *Encyclopedia of Human Factors and Ergonomics*.

Dumas, J.S. and Loring, B.A. (2008). *Moderating usability tests: Principles and practices for interacting*. Burlington: Morgan Kaufmann.

Dumas, J.S. and Redish, J. (1999). *A practical guide to usability testing*. Bristol: Intellect Books.

Dunn, O.J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241–252.

Ebling, M.R. and John, B.E. (2000). On the contributions of different empirical data in usability testing. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. New York: ACM. pp. 289–296.

Elbabour, F. (2015). *Eye Tracking in Retrospective Think Aloud Usability Testing: is there Added Value?* Unpublished: University of East Anglia. Master's dissertation.

Ericsson, K.A. and Simon, H.A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215.

Ericsson, K. A. and Simon, H.A. (1984). *Protocol Analysis: Verbal Reports* as *Data.* Cambridge: MIT Press.

Ericsson, K.A. (1988). Concurrent verbal reports on text comprehension: A review. *Text*, 4, 295–325.

Ericsson, K. A. and Simon, H.A., (1993) *Protocol Analysis: Verbal Reports as Data, Revised edition*. Cambridge: MIT Press.

Ericsson. K. A (2002). Towards a procedure for eliciting verbal expression of non–verbal experience without reactivity: interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology,*16(8), 981–987.

Ericsson. K. A. and Fox. M.e. (2011). Thinking aloud is not a form of introspection but a qualitatively different methodology: reply to Schooler (2011). Psychological Bulletin, 137(2), 351–354.

Faulkner, L. (2003). Beyond the five–user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, and Computers,* 35**,** 379–383.

Field, A. (2009). *Discovering statistics using SPSS, 2nd edition*. London: Sage.

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

Fox, M.C., Ericsson, K.A. and Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta–analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316.

Følstad, A. and Hornbaek, K. (2010). Work–domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing. *Journal of Systems and Software.* 83(11), 2019–2030.

Frøkjær, E., Hertzum, M., and Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* New York: ACM. pp. 345–352

Freeman, B. (2011). Triggered think–aloud protocol: using eye tracking to improve usability test moderation. In. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada. 07–12 May. New York: ACM. pp. 1171–1174.

Gable, G. (1994). Integrating Case Study and Survey Research Methods: an Example in Information Systems. *European Journal of Information Systems,* 3(2), 112–126.

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

Guan, Z., Lee, S., Cuddihy, E. and Ramey, J. (2006). The validity of the stimulated retrospective think–aloud method as measured by eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI '06. Montreal, Canada. 22–27 April. New York: ACM. pp. 1253–1262.

Guba, E. and Lincoln, Y. (1994). Competing Paradigms in Qualitative Research. In N. Denzin and Y. Lincoln (Ed.), *Handbook of Qualitative Research.* New York: Sage. pp. 105–117.

Gulliksen, J., Boivie, I. and Göransson, B. (2006). Usability professionals—current practices and future development. *Interacting with computers*, *18*(4), pp.568–600.

Hackos, J.T. and Redish, J. (1998). *User and task analysis for interface design.* Hoboken: Wiley Computer Publishing.

Hartson, H.R., Castillo, J.C., Kelso, J. and Neale, W.C. (1996). Remote evaluation: the network as an extension of the usability laboratory. In *Proceedings of the SIGCHI conference on human factors in computing systems*. New York: ACM. pp. 228–235

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), 373–410.

Hasan, L. (2009). *Usability evaluation framework for e–commerce websites in developing countries.* Unpublished: Loughborough University. PhD.

Hansen, M. (1991). Ten steps to usability testing. In *Proceedings of the 9th Annual International Conference on Systems Documentation*. New York: ACM. pp. 135–139

Hayes, J.R. and Flower, L.S. (1986). Writing research and the writer. *American Psychologist*, 41(10), 1106.

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), 421–443.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human–Computer Interaction*, 21(2), 125–146.

Hertzum, M., Hansen, K.D. and Andersen, H.H.K. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour and Information Technology*. 28(2), 165–181.

Hertzum, M., Molich, R. and Jacobsen, N.E. (2014). What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour and Information Technology*, 33(2), 144–162.

Hewett, T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., and Verplank, W. (1996). *ACM SIGCHI Curricula for Human–Computer Interaction.* New York: ACM.

Hirschheim, Rudy. (1985) "Information systems epistemology: An historical perspective." *Research Methods in Information Systems,* 13–35.

Holleran, P.A. (1991). A methodological note on pitfalls in usability testing. *Behaviour and Information Technology*, 10(5), 345–357

Hornbæk, K. and Frøkjær, E. (2006). What kinds of usability-problem description are useful to developers? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 50(24), 2523 –2527.

Hornbæk, K. and Law, E.L.C. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 617–626). ACM.

Hornbaek, K. and Frøkjær, E. (2008). A Study of the Evaluator Effect in Usability Testing. *Human–Computer Interaction*, 23(3), 251–277.

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

Hornbæk, K. (2011). Some whys and hows of experiments in human–computer interaction. *Human–Computer Interaction*, 5(4), 299–373.

Howarth, J., Smith–Jackson, T. and Hartson, R. (2009). Supporting novice usability practitioners with usability engineering tools. *International Journal of Human–Computer Studies*, 67(6), 533–549.

Howitt, D. and Cramer, D. (2007). *Introduction to Research Methods in Psychology*. Harlow: Prentice Hall.

Imler, B., and Eichelberger, M. (2014). Commercial database design vs. library terminology comprehension: Why do students print abstracts instead of full–text articles?. *College & Research Libraries*, 75(3), 284–297.

ISO 9241–11.2. (1996). *Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*.

ISO. (1999). 13407 Human–Centred Design Processes for Interactive Systems. Geneva.

Jacobsen, N.E., Hertzum, M. and John, B.E. (1998). The evaluator effect in usability tests. In *CHI 98 Conference Summary on Human Factors in Computing Systems*. New York: ACM. pp. 255–256

Jacobsen, Niels Ebbe. (1999). *Usability evaluation methods: the reliability and usage of cognitive walkthrough and usability test*. Unpublished Roskilde University. PhD.

Jeffries, R., Miller, J.R., Wharton, C. and Uyeda, K. (1991). User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM. pp. 119–124.

Jeng, J. (2005). What is usability in the context of the digital library and how can it be measured? *Information Technology and Libraries*, 24(2), 3–12.

Jensen, J.J. (2009). *Social Context in Usability Evaluations: Concepts, Processes and Products*. Unpublished: Aalborg University. PhD.

John, B.E. and Kieras, D.E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer–Human Interaction (TOCHI)*, *3*(4), 320–351.

Johnson, P. (1998). Usability and Mobility; Interactions on the move. In *Proceedings of the First Workshop on Human–Computer Interaction with Mobile Devices*.

Jørgensen, A.H., 1990. Thinking–aloud in user interface design: a method promoting cognitive ergonomics. *Ergonomics*, *33*(4), 501–507.

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

Karat, J. (1997). User–centred software evaluation methodologies. In M.G. Helander, T.K. Landauer and P.V. Prabhu (Eds.) *Handbook of human–computer interaction, 2nd edition*. Amsterdam: Elsevier Science. pp. 689–704.

Keenan, S.L., Hartson, H.R., Kafura, O.G., and Schulman, R.S. (1999). The usability problem taxonomy: a framework for classification and analysis. *Empirical Software Engineering*, 4(1). 71–104.

Kjeldskov, J. and Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human–Computer Studies*, 60(5), 599–620

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

Krathwohl, D. (1997). *Methods of Educational and Social Science Research: An Integrated Approach.* Boston: Addison Wesley Longman.

Kruskal, W.H. and Wallis, W.A. (1952). Use of ranks in one–criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.

Kumar, R. (2005). *Research Methodology: A Step–By–Step Guide for Beginners*. London: SAGE Publications.

Kumar, J., Yammiyavar, P. and Nielsen, J. (2008). Mind Tape technique–a usability evaluation method for tracing cognitive processes in cross cultural settings. *e–Minds, 1*, 69–85.

Lavery, D., Cockton, G. and Atkinson, M.P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour* and *Information Technology.* 16(4–5). 246–266.

Lazar, J., Feng, J.H., and Hochheiser, H. (2010). *Research Methods in Human–Computer Interaction.* Hoboken: Wiley.

Law, L.C. and Hvannberg, E.T. (2002). Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. In Proceedings of the Second Nordic Conference on Human–Computer Interaction. New York: ACM. pp. 71–80.

Law, E. L.C., and Hvanneberg, E. T. (2004). *Analysis of combinatorial user effects in international usability tests*. Unpublished paper presented at CONNECT: CHI2004. 24–29 April 2004, Vienna, Austria

Law, E.L. and Hvannberg, E.T. (2008) 'Consolidating usability problems with novice evaluators'. In *Proceedings of the 5th Nordic Conference on Human–Computer Interaction: Building Bridges.* NordiCHl'08. 18–22 October 2008, Lund, Sweden. New York: ACM. pp. 49–98.

Lindgaard, G. and Chattratichart, J., 2007, April. Usability testing: what have we overlooked?. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. New York: ACM. pp. 1415–1424.

Lewis, C. and Rieman, J. (1993). *Task–centered user interface design. A Practical Introduction.* New York: ACM.

Lewis, J., 2001. Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. *International Journal of Human–Computer Interaction*, 13(4), 445–79.

Lewis, J.R (2006a). Usability Testing. In: G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics.* Hoboken: Wiley. pp. 1275–1316.

Lewis, J.R. (2006b). Sample sizes for usability tests: mostly math, not magic. I*nteractions*, 13(6), 29–33.

Lewis, J.R. (2014). Usability: lessons learned… and yet to be learned. *International Journal of Human–Computer Interaction*, 30(9), 663–684.

Long, K. M., Styles, L. J., Andre, T. S., and Malcom, W.C. (2005). Usefulness of nonverbal cues from participants in usability testing sessions. In G. Salvendy (Ed.),

Proceedings of the Human–Computer–Interaction International Conference (CD ROM Vol. 4). St. Louis, MO: Mira Digital Publishing.

Macefield, R. (2007). Usability studies and the Hawthorne Effect. *Journal of Usability Studies*, 2(3), 145–154.

Macefield, R., 2009. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies*, 5(1), 34–45.

Makri, S., Blandford, A., and Cox, A. L. (2011). This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behaviour. *Information Processing and Management*, 47(3), 336–348.

Maguire, M. (2001). Context of Use within usability activities. *International Journal of Human–Computer Studies.* 55(4), 453–483.

Marshall, C. and Rossman, G.B. (1999). Defending the value and logic of qualitative research. *Designing Qualitative Research*, 191–203.

Martin, R., Al Shamari, M., Seliaman, Mohamed E., Mayhew, P. (2014). Remote Asynchronous Testing: A Cost–Effective Alternative for Website Usability Evaluation. *International Journal of Computer and Information Technology*, 3(1), 99–104.

Maxwell, J. (2005). *Qualitative Research Design: An Interactive Approach*. Beverley Hills: Sage Publications.

McCambridge, Jim, Witton, John and Elbourne, Diana R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67(3), 267–277.

McDonald, S., Monahan, K. and Cockton, G. (2006). Modified contextual design as a field evaluation method. In *Proceedings of the 4th Nordic conference on Human–Computer Interaction: Changing Roles*. New York: ACM. pp. 437–440

McDonald, S., Edwards, H. and Zhao, T. (2012) Exploring think-alouds in usability testing: an international survey. *IEEE Transactions on Professional Communication*, 55(1), 1–17.

McDonald, S., Zhao, T. and Edwards, H.M. (2013). Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human–Computer Interaction*, 29(10), 647–660.

McNamara, N. and Kirakowski, J. (2005). Defining usability: quality of use or quality of experience? In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005*. New York: IEEE. pp. 200–204

Meissner, C.A. and Brigham, J.C. (2001). A meta–analysis of the verbal overshadowing effect in face identification. Applied Cognitive Psychology. 15(6). 603–616.

Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B. (2004). Comparative usability evaluation. *Behaviour and Information Technology*, 23(1), 65–74.

Molich, R and Dumas, J. (2008) Comparative usability evaluation (CUE–4). *Behaviour* and *Information Technology,* 27(3). 263–281.

Monahan, K., Lahteenmaki, M., McDonald, S. and Cockton, G. (2008). An investigation into the use of field methods in the design and evaluation of interactive systems. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, Volume 1*. Swindon: British Computer Society. pp. 99–108

Munzner, T., 2008. Process and pitfalls in writing information visualization research papers. In *Information visualization*. Heidelberg: Springer. pp. 134–153

Naveedh, S. (2015). *Exploring the Use of Think-Aloud Methods in Usability Practice.* Unpublished: University of East Anglia. Master's dissertation.

Newman, W. L. (2003). *Social research methods: qualitative and quantitative approach.* Sydney: Pearson Education, Inc.

Nielsen, J., (1992). Finding usability problems through heuristic evaluation. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 373–380.

Nielsen J (1993a). *Usability Engineering. San Francisco:* Morgan Kaufmann Press.

Nielsen, J. and Landauer, T.K. (1993b). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human Factors in Computing Systems*. New York: ACM. pp. 206–213.

Nielsen, J., (1994a). Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier, in In Randolph G. Bias, Deborah J. Mayhew (Eds.), *Cost–Justifying Usability.* Burlington: Morgan Kaufmann Press. pp. 245–272.

Nielsen, J. and Levy, J. (1994b). Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4), 66–75.

Nielsen, J. (1995). 10 Usability Heuristics for User Interface Design. [Online] NN Group. Available at: https://www.nngroup.com/articles/ten-usability-heuristics/

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

Nielsen, J., (2000). Why You Only Need to Test with 5 Users. [Online] NN Group Available at https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users

Nielsen, J., (2002), Becoming a Usability Professional. [Online] NN Group. Available at: https://www.nngroup.com/articles/becoming-a-usability-professional/

Nielsen, J., and Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4), 66–75.

Nielsen, J. (2010). Testing expert users. [Online] *Jakob Nielsen's Alertbox.* Available: https://www.nngroup.com/articles/testing-expert-users/

Norman, K.L. and Panizzi, E (2006). Levels of automation and user participation in usability testing. *Interacting with Computers.* 18(2), 246–264.

Nørgaard, M. and Hornbaek, K. (2006) What do usability evaluators do in practice?: an explorative study of think–aloud testing'. In: *Proceedings of the fifth Conference on Transfer Designing Interactive Systems*, University Park, USA, 26–28 June 2006. New York: ACM. pp. 209–218

Nørgaard, M. and Hornbaek, K. (2008). Working together to improve usability: Challenges and Best Practices. In: *Copenhagen University Technical Report.* University of Copenhagen. [Online] Available at: www.diku.dklOLO/publikationerltekniske.rapporter/rapporter/O8–03.pdf.

Ohnemus, K.R. and Biers, D.W. (1993). Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting Vol. 37, No. 17*. New York: SAGE Publications. pp. 1127–1131.

Olmsted–Hawala, E.L., Murphy, E.D., Hawala, S. and Ashenfelter, K.T. (2010). Think–aloud protocols: a comparison of three think–aloud protocols for use in testing data–dissemination web sites for usability. In *Proceedings of the SIGCHI conference on human factors in computing systems*. New York: ACM. pp. 2381–2390.

Olmsted–Hawala, E. and Bergstrom, J.R. (2012). Think–aloud protocols: does age make a difference? *Proceedings of Society for Technical Communication (STC) Summit, Chicago, IL*.

Oostendorp, van, H., and De Mul, S. (1999). Learning by exploration: thinking aloud while exploring an information system. *Instructional Science*, 27, 269–284.

O'Malley, C., Baker, M., and Elsom–Cook, M. (1991). The design and evaluation of a multimedia authoring system. *Computers & Education*, 17, 49–60.

O'Rourke, N. and Hatcher, L. (2008) *A Step–by–Step Approach to Using SAS for Univariate and Multivariate Statistics*. London: SAS Publishing.

Perfetti, C., and Landesman, L. (2002). Eight is not enough. [Online] Available at: http://world.std.com/~uieweb/Articles/eight_is_not_enough.htm.

Peute, L.W., de Keizer, N. F. and Jaspers, M.W.M., (2010) Cognitive evaluation of a physician data query tool for a national ICU registry: comparing two think aloud variants and their application in redesign. *Studies in Health Technology and Informatics*, 160(1), 309–313.

Pyzdek, T. (2003). *The six sigma handbook, 2nd ed.*New York: McGraw-Hill

Robson, C. (2002). *Real world research: A Resource for Social Scientists and Practitioner-Researchers. 2nd Ed.* Oxford: Blackwell.

Rogers, Y. (2011). Interaction design gone wild: striving for wild theory. *Interactions*, 18(4), 58–62.

Rubin, J. and Chisnell, D. (2008). *Handbook of usability testing: how to plan, design and conduct effective tests*. Hoboken: Wiley.

Rubin, A., and Babbie, E. R. (2009). Essential research methods for social work: Brooks. *Cole Pub Co*.

Rudd, J., Stern, K. and Isensee, S. (1996). Low vs. high–fidelity prototyping debate. *Interactions*, 3(1), 76–85.

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), 670–677.

Sauer, J., Seibel, K. and Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied ergonomics*, 41(1), 130–140.

Saunders, M., Lewis, P. and Thornhill, A. (2007). *Research methods for business students* (4th Ed.). Essex: Pearson Education.

Sauro, J. (2010). A Practical Guide to Measuring Usability: 72 Answers to the Most Common Questions about Quantifying the Usability of Websites and Software. A Measuring Usability LLC.

Sauro, J. and Lewis , R. (2012).Quantifying The User Experience: Practical Statistics For User Research. Elsevier.

Sauro, Jeff. (2013). *17 Periodicals for Usability Research*. [Online] Available at: http://www.measuringu.com/blog/usability-periodicals.php

Sauro, J. (2015). How Confident Do You Need to be in Your Research?, [Online] Available at: http://www.measuringu.com/blog/confidence-levels.php

Scholtz, J. (2006) Usability evaluation. National Institute of Standards and Technology. *Human–computer interaction: new trends; 13th international conference.* July 2009. San Diego, USA. New York: ACM Press.

Schooler, J.W. and Fiore, S.M. (1997). Consciousness and the limits of language: You can't always say what you think or think what you say. In: J. D. Cohen and J. W. Schooler (eds.). *Scientific Approaches to Consciousness.* New Jersey: Lawrence Erlbaum Associates, pp. 241–293

Schooler, J.W. (2011). Introspecting in the spirit of William James: comment on Fox, Ericsson, and Best (2011). Psychological Bulletin. 137 (2). pp. 345-350.

Sekaran, U. and Bougie, R. (1992). Business Research Methods.

Séguinot, C. (1996). Some Thoughts About Think–Aloud Protocols. *Target*, *8*, 75–95.

Shi, Q. (2008). A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In *Proceedings of the 5th Nordic Conference on Human–Computer Interaction: Building Bridges.* NordiCHI'08. 18–22 October 2008, Lund, Sweden. New York: ACM. pp. 344–352.

Shi, Q. (2010). *An empirical study of thinking aloud usability testing from a cultural perspective: PhD Series 30.2010*. Copenhagen: Copenhagen Business School.

Simons, H. (2010). *Case Study Research in Practice.* London: SAGE Publications.

Skov, M.B. and Stage, J. (2012). Training software developers and designers to conduct usability evaluations. *Behaviour and Information Technology*, 31(4), 425–435.

Sova, D.H., Nielsen, J. and NN GROUP (2003). 234 Tips and Tricks for Recruiting Users as Participants in Usability Studies. [Online] NN Group. Available at: http://www.nngroup.com/reports/tips/recruiting/234_recruiting_tips.

Sonderegger, A., Schmutz, S. and Sauer, J. (2016). The influence of age in usability testing. *Applied Ergonomics*, *52*, 291–300.

Sonderegger, A. (2010). *Influencing factors in usability tests*. Unpublished: Université de Fribourg. PhD.

Spool, J. and Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. In *CHI'01: Extended Abstracts on Human Factors in Computing Systems*. New York: ACM. pp. 285–286.

Sun, X., and Shi, Q. (2007). Language issues in cross cultural usability testing: A pilot study in China. Paper presented at the HCI International Conference 2007, Beijing.

Thomas, Nathan (2015). How To Use The System Usability Scale (SUS) To Evaluate The Usability Of Your Website. [Online] Usability Geek. Available at: http://usabilitygeek.com/how-to-use-the-system-usability-scale-sus-to-evaluate-the-usability-of-your-website. Accessed: [insert date here].

Tullis, T. and Albert, B. (2008). *Measuring the user experience*. Burlington: Elsevier Inc.

Turner, C. W., Lewis, J. R., and Nielsen, J. (2006). Determining usability test sample size. *International Encyclopedia of Ergonomics and Human Factors*, 3(2), 3084–3088.

Van den Haak, M.J., de Jong, M.D. and Schellens, P.J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with computers*, 16(6), 1153–1170.

Vermeeren, A. P. O. S., Bouwmeester, K., Aasman, J. and de Ridder, H. (2002). DEVAN: A tool for detailed video analysis of user test data. *Behaviour and Information Technology*, 21(6), 403–423.

Vermeeren, A. P.O.S, van Kesteren, I.E., and Bekker, M. M (2003). Managing the evaluator effect in user testing. In *Proceedings of the 2003 IFIP TC13 Intemational Conferences on Human–Computer Interaction*. INTERACT'03. Amsterdam: lOS Press, 647—654.

Virzi, R., (1992). Refining the test phase for usability evaluation: How many subjects is enough?, *Human Factors*, 457–86

Virzi, R.A., Sokolov, J.L. and Karis, D. (1996). Usability problem identification using both low–and high–fidelity prototypes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM. pp. 236–243.

Watson, J.B. and Rayner, R., (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, 3(1),1.

Wichansky, A.M. (2000). Usability testing in 2000 and beyond. *Ergonomics*, 43(7), 998–1006.

Wilson, T.D., (1994). Commentary to feature review: the proper protocol: validity and completeness of verbal reports. 249-252.

Wilson, T.D. (2004). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge: Belknap Press/Harvard University Press.

Wixon, O. (2003) 'Evaluating usability methods: why the current literature fails the practitioner.' *Interactions.* 10(4). 28–34.

Willis, L.M. and McDonald, S. (2016). Retrospective protocols in usability testing: a comparison of Post–session RTA versus Post–task RTA reports. *Behaviour and Information Technology*, 1–16.

Wisker, G. (2001). *The Postgraduate Research Handbook*. Basingstoke: Palgrave.

Woolrych, A. and Cockton, G. (2001). Why and when five test users aren't enough. In Eds. *Proceedings of IHM–HCI 2001 conference*, Vol. 2,. Toulouse: Cépaduès Editions. pp. 105–108

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), 940–970.

Wright, P.C. and Monk, A.F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man–Machine Studies*, 35(6), 891–912.

Yin, R.K. (1984). *Case Study Research: Design and Methods*. Beverly Hills: Sage Publications.

Zhao, T., and McDonald, S. (2010). Keep talking: an analysis of participant utterances gathered using two concurrent think–aloud methods. In *Proceedings of the 6th Nordic Conference on Human–Computer Interaction: Extending Boundaries*. New York: ACM. pp. 581–590.

Zhao, T., McDonald, S. and Edwards, H.M. (2012). The impact of two different think–aloud instructions in a usability test: a case of just following orders?. *Behaviour and Information Technology*, 33(2), 163–183.

APPENDICES

Appendix A: Usability Heuristic Evaluation Checklist

1. Visibility of system status The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. Match between system and the real world The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
3. User control and freedom Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
4. Consistency and standards Users should not have to wonder whether different words, situations, or actions mean the same thing.
5. Error prevention Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
6. Recognition rather than recall Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. Flexibility and efficiency of use Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8. Aesthetic and minimalist design Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
9. Help users recognize, diagnose, and recover from errors Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
10. Help and documentation Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Adopted from (Nielsen, 1995)

Appendix B: Research Design

Appendix B1: Experience with TA Test Questionnaire

1. Participant ID

2. Working Conditions:
To what extent do you agree or disagree with the following statements about your working procedure?

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
1. The working procedure on the tasks was slower than my normal working procedure.	○	○	○	○	○
2. The working procedure on the tasks was more focused than my normal working procedure.	○	○	○	○	○

3. Thinking Aloud Experience:
To what extent do you agree or disagree with the following statements about your think-aloud experience ?

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
1. Thinking-aloud is difficult	○	○	○	○	○
2. Thinking-aloud is unnatural	○	○	○	○	○
3. Thinking-aloud is unpleasant	○	○	○	○	○
4. Thinking-aloud is tiring	○	○	○	○	○
5. Thinking-aloud is time-consuming	○	○	○	○	○

4. Evaluator Presence:
To what extent do you agree or disagree with the following statements about the presence of the evaluator?

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
1. I felt it was unnatural to have the evaluator present during the study.	○	○	○	○	○
2. It was disturbing to have the evaluator present during the study.	○	○	○	○	○
3. It was unpleasant to have the evaluator present during the study.	○	○	○	○	○

Appendix B2: System Usability Scale Questionnaire

*** 1. ID Number:**

*** 2. After Using the website, please indicate your opinion accordingly.**

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

Appendix B3: Problem Indicators Checklist

Definition	Description
Indications types based on verbal and/or non-verbal behaviour	
Puzzled	User indicates: Uncertainty about what actions to take To be sure whether a specific action is needed or not Not being able to understand something on the system (e.g. informative text, a link name, terminology, or a function).
Wrong explanation or Understanding	The user gives an explanation of something that has happened but this explanation is incorrect User verbalises an incorrect understanding of something on the system (e.g. informative text, a link name or functionality)
Recognition	User indicates they recognise a preceding error User indicates that they now understand something previously not understood.
Design suggestion	User makes a design suggestion
Quit Task	The user declares that they are abandoning a task The user recognises that the current task was not finished successfully, but continues with a subsequent task
Doubt, Surprise, Frustration	The user indicates: They are unsure as to where they have and have not been on the system They are unsure as to whether an action has executed properly Do not understand an actions effect To be surprised by an action's effect That something did not meet their expectations The effect of an action was unsatisfactory or frustrated the user They dislike or disapprove of something
Random Actions	The user indicates verbally or non-verbally that they are performing random actions.
Impatience	The user shows impatience by clicking repeatedly on objects that respond slowly or the user expressed impatience verbally.
Wrong goal	User formulates a goal that cannot be achieved
Search for function	User indicates Not being able to locate a specific functional link or piece of information They are searching for a function the evaluator knows does not exist.
Indication types based on observed actions:	
Wrong Action	User points at a correct function/object but does not execute the action Execution of an action not done correctly or optimally User stops executing a correct action before it is finished An action does not belong to the correct sequence of actions An action is omitted from the sequence An action within a sequence is replaced by another action Actions within a sequence are performed in reverse order
Repeated Action	User has to re-do certain actions (e.g. re-enter form data due to it not being saved) User repeats an action with the same effect
Technical Issues	System crashes, broken links, slow response system

Appendix B4: Indivdual Problem Report

Individual Problem Report		
IUP No.:	Participant's No.:	Task#:
Problem indicator#:		Timestamp:
Source:		
Context (the user's goal):		
Usability problem (the user's difficulty and associated causes):		
Impact:		
Severity:	Persistence:	

Appendix B5: Final Problem Report

Final Problem Report

FUP No.: Frequency:

Context (the user's goal):

Brief description of the final problem:

Associated IUPs (all the IUPs that form this final problem)

(Participant NO.-Individual problem number-Severity level-Source)

Final Severity rating: Final Problem source: Probelm Type:

Appendix C: Materials from Study One

Appendix C1: UEA Approval

Dear Obead,
The submission of your proposal has been considered by the UEA Computing Sciences Research Ethics Committee and I can confirm that your proposal has been approved.

Please could you ensure that any further amendments to either the protocol or documents submitted are notified to me, as Chair of CMP-REC, in advance and also that any adverse events which occur during your project are reported to the Committee.

The Committee would like to wish you good luck with your project

Best wishes,
Dan Smith
(Chair CMP-REC)

Dr D.J. Smith email: Dan.Smith@uea.ac.uk
School of Computing Sciences tel: +44 (0)1603 592608
University of East Anglia fax: +44 (0)1603 593345
NORWICH NR4 7TJ, UK

Appendix C2: Email Sent to the Administrator of the Website

Dear Sir or Madam,

Thank you for your interest in my email.

My name is Obead Alhadreti, and I am a PhD student in human-computer interaction in the School of Computing Sciences at the University of East Anglia.

I am writing to you to seek your kind approval to use your University library's website as a test object in my usability study.

Sixty participants will be recruited to conduct the experiment applying think-aloud methods. At the end of the study, I will provide you with a report that lists all usability problems found on your website.

I look forward to hearing from you at your earliest convenience.

Best regards,

Obead Alhadreti

Appendix C3: Website's Administrator Approval

Hi Obead,

I'd be happy for you to undertake this project on the library website, and I'd definitely be interested in hearing the results, I should mention that our website content management system is managed by our IT department, so unfortunately adding in some of the accessibility features that I would ideally put in place is the sort of decision they have to make. However, the more information and evidences we have about this important area, the more likely we are to be able to persuade them so I'm all for it.

With this in mind, if are still interested in using the library website for your project do go ahead. I would definitely be interested in hearing your findings.

Best wishes,

Russel

Appendix C4: Interview Agenda

Interview Agenda		
Location: Telephone	**Date:** 02-07-2013	**Time:** 12:00-12:30
Aim of the interview: To understand the intended audience of the library website and their activities on the site.	**Interviewee's rights:** Taking part is entirely voluntary. You are free to withdraw at any time without any penalty and the data will be destroyed.	**The interviewer role:** As the interviewer, I will be taking notes and will be recording your voice for later analysis.
Type of interview: structured. A list of interviewing questions is prepared.	**Data confidentiality:** All collected data will be kept strictly confidential and will not be made public in any way.	**Publication:** The results of the analysis of this interview may be published, but you as an individual will not be identifiable. If you would like to access to any reports or publications, please let me know
Permission to audio record: With your consent, I would like to audio record the interview. This will allow me to focus completely on what you are saying. No one outside the research team will have access to the audio recording or to any information that could identify you. The audio recording will be deleted at the end of the project.		

Appendix C5: Task List

Task ID	Task Description
T1	You are encountering difficulty in finding a specific book on the subject of computing that you need to read before the exams. Using the website, please find the name of an academic support librarian for the subject of computing. *Can you find it?*
T2	You want to book a room at the library to study for your coming exam. Using the website, find the next available time for study rooms. *Can you find it?*
T3	You are a big fan of the author "Austin Sarat" and want to know how many publications are written by your favourite author. *Can you find it?*
T4	You want to find the journal paper that has the title "Building for the Future" written by Doyle Henry in 1963 to read before coming seminar in education subject. Can you find it?
T5	You want to find how many books that have the keywords "climate change" in their titles were published in the last five years. *Can you find them?*
T6	You want to find the citation for the book '*Mobile Usability'* to add it to the paper that you are writing. *Can you find it?*
T7	You want to view your previous search history for academic resources on the website so you can remember the titles of some important resources that you looked for before. *Can you find the previous search history?*

Appendix C6: Screening Questionnaire

* 1. Please enter your contact information below:

Name (optional):

Email address:

* 2. Your current status is:

- UEA student
- UEA academic staff

Other (please specify)

* 3. Which category below includes your age?

- Younger than 18
- 18-28
- 29-39
- 40-50
- 51-65
- More than 65

* 4. What is your gender?

- Male
- Female

* 5. What is your nationality?

6. Is English your first language?

- Yes
- NO

* 7. If English is not your mother tongue, please check your latest overall IELTS score (or any equivalent):

- Less than 6.5
- + 6.5

* 8. Please check your current educational level below:

- Undergraduate student
- Postgraduate student (Master, MPhil, or PhD)
- Other (please specify)

9. Do you consider yourself to have any of the following:

- Serious hearing, visual or speech impairments
- Need assistive technologies to use a computer
- Social/communication impairment
- Mental or learning difficulties
- Other serious disability or impairment that is not listed
- Prefer not to say
- No, I do not have any

* 10. How long have you been using the Internet, not including time spent working with email?

- ◯ More than 5 years
- ◯ More than 3 years
- ◯ More than 1 year
- ◯ Less than a year

* 11. How often do you use the Internet, not including time spent working with email?

- ◯ Every day
- ◯ At least once a week
- ◯ A few times a month
- ◯ Less than once a month

* 12. Have you used any online university library before?

- ◯ No
- ◯ Yes, if 'yes' please state which library website (s) you used starting with most recent ones:

* 13. Have you participated in usability evaluation before?

- ◯ Yes
- ◯ No

14. What Internet browser do you usually use?

- ◯ Internet Explorer
- ◯ Google Chrome
- ◯ Firefox
- ◯ Safari
- ◯ Opera
- ◯ Maxthon
- ◯ Don't Know
- ◯ Other (please specify)

* 15. Are you willing to have your voice, face, and on-screen computer actions recorded during the session for analysis purposes only? Your information will be kept confidential.

- ◯ Yes
- ◯ No

Thank you for taking the time to complete this survey. If you are selected to participate in this study, you will be contacted in two to three weeks with further information. Please click 'Done' to submit your answers. Thanks again.

Appendix C7: Email Sent to Students

Participants Needed!

What an Easy Way to Get £5

Hello,

My name is Obead Alhadreti, 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 be asked to try out a website by performing a few activities on your own, and to give me your feedback. You will also fill in a short questionnaire about your experience with the session.

When and where?

The study will be conducted in the school of Computing Sciences at the University of East Anglia from the 15th of October until the 5th of December 2013.

Why to get involved?

- ✓ **Financial reward:** If selected to participate, you will receive £5 as token of appreciation.
- ✓ **Confidentiality:** All data will be kept confidential and treated anonymously.
- ✓ **Short time:** The study should take at most no more 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-minute screening survey:

Click here to take part.

The survey will close on Monday 15 September. If you meet the criteria I am seeking for the purposes of this research, you will be contacted by email with further information regarding the study.

Your contribution is highly appreciated. If you would like more information, please contact me at:

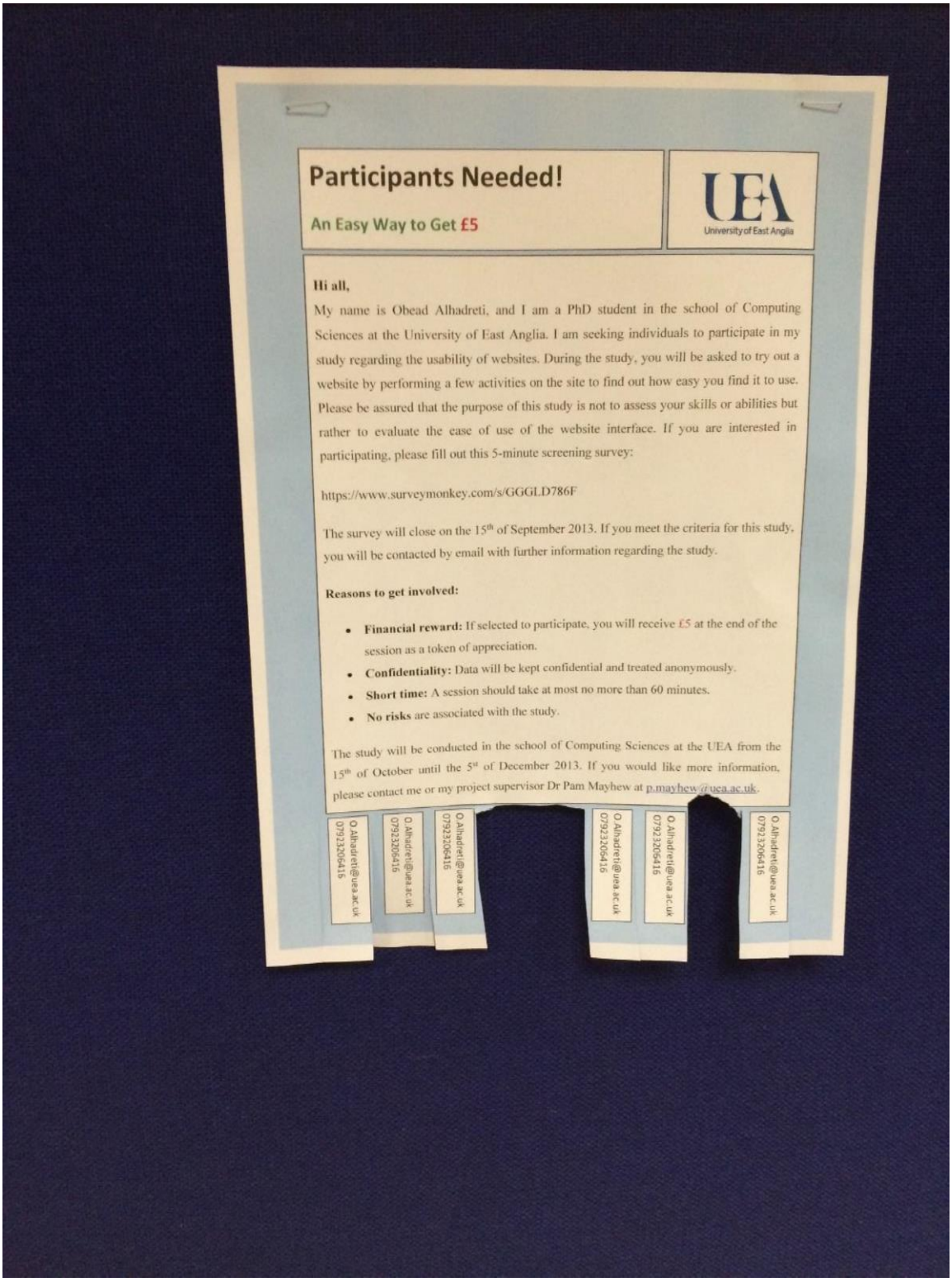
O.Alhadreti@uea.ac.uk, 07923206416

Thank you for your interest,

Obead Alhadreti

Researcher
School of Computing Sciences,
UEA, UK

Appendix C8: Poster Displayed to Students



Participants Needed!

An Easy Way to Get £5

University of East Anglia

Hi all,

My name is Obead Alhadreti, 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 my study regarding the usability of websites. During the study, you will be asked to try out a website by performing a few activities on the site to find out how easy you find it to use. Please be assured that the purpose of this study is not to assess your skills or abilities but rather to evaluate the ease of use of the website interface. If you are interested in participating, please fill out this 5-minute screening survey:

https://www.surveymonkey.com/s/GGGLD786F

The survey will close on the 15th of September 2013. If you meet the criteria for this study, you will be contacted by email with further information regarding the study.

Reasons to get involved:

- **Financial reward:** If selected to participate, you will receive £5 at the end of the session as a token of appreciation.
- **Confidentiality:** Data will be kept confidential and treated anonymously.
- **Short time:** A session should take at most no more than 60 minutes.
- **No risks** are associated with the study.

The study will be conducted in the school of Computing Sciences at the UEA from the 15th of October until the 5st of December 2013. If you would like more information, please contact me or my project supervisor Dr Pam Mayhew at p.mayhew@uea.ac.uk.

O.Alhadreti@uea.ac.uk 07932306416

Appendix C9: Invitation Email Sent to Students

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 mystery 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 recoding your voice and capturing your actions on the computer screen; however, these recordings will be for research purposes only and will not be made public in any way. At the completion of the website interaction, you will be asked to complete a short online questionnaire regarding your experience with the session. 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: 2.17 in the School of Computing Sciences at the University of East Anglia. The whole session should take at the most 60 minutes, depending on your level of comfort. At the end of your session, you will receive 5 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.

http://doodle.com/polls/notifications?participantId=1956741959&pollId=nz4833xnfgwibq5

For further information about the study location, please click on the following link:

http://doodle.com/polls/notifications?participantId=1956741959&pollId=nz4833xnfgwibq5

I will send you a reminder email a couple of days before your session. Thank you for agreeing to participate in my study and for making the web *a better place*.

Sincerely,

Obead Alhadreti

Appendix C10: Confirmation Email Sent to Students

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: 2.17 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,

Obead Alhadreti

O.Alhadreti@uea.ac.uk

Appendix C11: Experiment Checklist

Part A. Before each experiment

- Ensure lab environment is comfortable
- Make copies of all study materials (pre-experiment questionnaire, procedure instruction sheet, consent form, receipt form, tasks sheet and list, observation sheet)
- Ensure lab and data recording equipment is running properly
- Make sure incentives for participants are ready
- Turn off or disable anything on the test computer that might interrupt the test (e.g., email or instant messaging, scheduled virus scans)
- Open the website home page and make sure the site is running properly
- Clear the browser history
- Create new folder for the test
- Turn off the participants and my mobile
- Get a glass of water
- Put 'Do not Disturb' sign on the door

Part B. Before each task

- When needed, reminder participant to go to home page of the website

Part C. During the experiment

- Where appropriate, encourage participants to think aloud if they stop for 15 seconds

Part D. At the end of each experiment

- End session recordings
- Save the recording
- Give incentive to participant
- Explain the real aim of the study to participants and justify why they had not been informed about it
- Answer any questions they may have
- Thank them and escort them out
- Post the video recording to predetermined location
- Back up all video and data files

Appendix C12: Consent Form

Consent Form

(Please read and sign this form)

Thank you for agreeing to participate in this study. The aim of this study is to evaluate a university library website. You will be encouraged to share your thoughts by thinking aloud (I will explain later what do I mean by this). During the study, it will be necessary for me to record a number of things using screen capture software, video and audio. However, this recorded data will be stored securely on a password-protected computer in accordance to the UEA's data protection policy. You can withdraw from this study at any time. Recordings and notes taken will be destroyed as you require.

Please tick the box for things that you agree with, and sign below if you are happy to give your consent for the study to go ahead.

1. Your monitor and voice will be recorded using screen capture. [ ]
2. The results of the analysis of this evaluation may be published, but all the data recorded will be anonymous. [ ]

* If you would like to access to any reports or publications that directly result from your involvement in this study, please tick the box. [ ]

Participant Name	*Signature of Participant*	*Date*
_____	_____	___ / ___ /201

Contact details:

Researcher:	Obead Alhadreti	Supervisor: Dr. Pam Mayhew
Email address:	O.Alhadreti@uea.ac.uk	P.Mayhew@uea.ac.uk
Contact number:	07923206416	01603593334

Appendix C13: CTA Condition Procedure Sheet

Hi and thank you for coming today. As you probably know my name is 'Obead'.

1. Please review and sign the informed consent form which will provide you with an overview of the study.

2. Please take a moment to familiarize yourself with the laptop and Internet browser.

3. Thinking Aloud: 'In this study, I am interested in what you say to yourself as you perform some tasks that I give you. In order to do this I will ask you to think aloud as you work on the tasks. What I mean by think-aloud is that I want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?'

4. Let's take a moment to practice thinking aloud. Please think aloud as you look up for the word 'carol' in the online dictionary opened in the browser. Please note that this is not the website under evaluation.
   Do you have any questions about the thinking aloud process you've just practiced?

5. Please read carefully task instructions on the screen.

6. If you don't have any questions, please start performing the first task written on the note card, placed on the table on your right hand, using the website. Please remember to think aloud while you are solving the tasks from beginning till the end of the task.

7. Now you have finished the tasks, please answer the online post-experiment questionnaires.

Thank you very much for your time and input. Here is your reward. Have a great day!

Appendix C14: RTA Condition Procedure Sheet

Hi and thank you for coming today. As you probably know my name is 'Obead'.

1. Please fill in this short background questionnaire.

2. Please now review and sign the informed consent form, which will provide you with an overview of the study.

3. Please take a moment to familiarize yourself with the laptop and Internet browser.

4. As a warm-up task, please look up for the word 'carol' in the online dictionary opened in the browser. Please note that this is not the website under evaluation.

5. Please read carefully the task instructions sheet on the next page.

6. If you don't have any questions, please start performing the first task written on the note card, placed on the table on your right hand, using the website. I would like you to solve the tasks in silence, just as if you were using the site at home.

7. Now you have finished the tasks, please complete the first two parts of the online post-experiment questionnaire.

8. 'I would like you now to please watch your recorded tasks performance on muted video and give retrospective reporting on them. In other words, I would like you to recall the thoughts you had when completing each task, and tell me any thoughts you had. If you have any questions, please ask them now'

9. As a practice task, I will show you now what you did when you were performing the warm-up task.

10. Now you have finished, the please complete the remaining two parts of the online post-test questionnaire.

Thank you very much for your time and input. Here is your reward. Have a great day!

Appendix C15: HB Condition Procedure Sheet

Hi and thank you for coming today. As you probably know my name is 'Obead'.

11. Please fill in this short background questionnaire.

12. Please now review and sign the informed consent form which will provide you with an overview of the study.

13. Please take a moment to familiarize yourself with the laptop and Internet browser.

14. Thinking Aloud: 'In this study, I am interested in what you say to yourself as you perform some tasks that I give you. In order to do this I will ask you to think aloud as you work on the tasks. What I mean by think-aloud is that I want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?'

15. Let's take a moment to practice thinking aloud. Please think aloud as you look up for the word "carol" in the online dictionary opened in the browser. Please note that this is not the website under evaluation.

    Do you have any questions about the thinking aloud process you've just practiced?

16. Please read carefully task instructions on the third page.

17. If you don't have any questions, please start performing the first task written on the note card, placed on the table on your right hand, using the website. Please remember to think aloud while you are solving the tasks from beginning till the end of the task.

18. Now you have finished the tasks. Please complete the first three parts of the online post-experiment questionnaire.

19. 'I would like you now to watch your recorded tasks performance on muted video and give retrospective reporting on them. In other words, I would like you to recall the thoughts you had when completing each task, and tell me any thoughts you had. If you have any questions, please ask them now; if not, you may begin'.

20. Now you have finished, please complete the remaining two parts of the online post-test questionnaire.

Thank you very much for your time and input. Here is your reward. Have a great day!

Appendix C16: Task Instructions Sheet

You have seven tasks to perform on University of East London library's website, each task is written on a separate note card.

- Please perform the tasks in the order presented, solve one task at a time, and make sure you understand each task requirements fully before you start. Feel free to ask questions if you are not sure about the task requirements.
- The website's homepage contains a major search feature, please use this only if they felt they had no other choice to solve a task.
- When you are going to start a task, please verbally alert me.
- Once you start the task, please try to solve it yourself, just like when you are using the website at home. I will not be able to offer any suggestions or hints.
- When you think that you have found the information you have been looking for please state 'your answer' out loud. If you feel you are unable to complete the task and would like to stop, please say 'moving on to next task' so I know and proceed to the next task.
- At times, I may ask you to move on to the next task even though you haven't finished the current task; this just because that I have obtained all the information that I needed.

Please ask if you are unsure about any of these instructions or have any questions at all. If you don't have any questions, please start the first task displayed in the note card labelled 'Task 1'.

Appendix C17: Task Counter Balancing

Methods	Participants	Order of task presentation						
CTA	**P1**	1	2	3	4	5	6	7
	P2	2	3	4	5	6	7	1
	P3	3	4	5	6	7	1	2
	P4	4	5	6	7	1	2	3
	P5	5	6	7	1	2	3	4
	P6	6	7	1	2	3	4	5
	P7	7	1	2	3	4	5	6
	P8	1	2	3	4	5	6	7
	P9	2	3	4	5	6	7	1
	P10	3	4	5	6	7	1	2
	P11	4	5	6	7	1	2	3
	P12	5	6	7	1	2	3	4
	P13	6	7	1	2	3	4	5
	P14	7	1	2	3	4	5	6
	P15	1	2	3	4	5	6	7
	P16	2	3	4	5	6	7	1
	P17	3	4	5	6	7	1	2
	P18	4	5	6	7	1	2	3
	P19	5	6	7	1	2	3	4
	P20	6	7	1	2	3	4	5
RTA	**P21**	1	2	3	4	5	6	7
	P22	2	3	4	5	6	7	1
	P23	3	4	5	6	7	1	2
	P24	4	5	6	7	1	2	3
	P25	5	6	7	1	2	3	4
	P26	6	7	1	2	3	4	5
	P27	7	1	2	3	4	5	6
	P28	1	2	3	4	5	6	7
	P29	2	3	4	5	6	7	1
	P30	3	4	5	6	7	1	2
	P31	4	5	6	7	1	2	3
	P32	5	6	7	1	2	3	4
	P33	6	7	1	2	3	4	5
	P34	7	1	2	3	4	5	6
	P35	1	2	3	4	5	6	7
	P36	2	3	4	5	6	7	1
	P37	3	4	5	6	7	1	2
	P38	4	5	6	7	1	2	3
	P39	5	6	7	1	2	3	4
	P40	6	7	1	2	3	4	5
HB	**P41**	1	2	3	4	5	6	7
	P42	2	3	4	5	6	7	1
	P43	3	4	5	6	7	1	2
	P44	4	5	6	7	1	2	3
	P45	5	6	7	1	2	3	4
	P46	6	7	1	2	3	4	5
	P47	7	1	2	3	4	5	6
	P48	1	2	3	4	5	6	7
	P49	2	3	4	5	6	7	1
	P50	3	4	5	6	7	1	2
	P51	4	5	6	7	1	2	3
	P52	5	6	7	1	2	3	4
	P53	6	7	1	2	3	4	5
	P54	7	1	2	3	4	5	6
	P55	1	2	3	4	5	6	7
	P56	2	3	4	5	6	7	1
	P57	3	4	5	6	7	1	2
	P58	4	5	6	7	1	2	3
	P59	5	6	7	1	2	3	4
	P60	6	7	1	2	3	4	5

Appendix C18: Observation Sheet

Usability Test Observation Sheet

Participant #: ____	**TA method:** ____	**Date:** / /2013
Session starts at: ___ h ___m		**ends at:** ___ h ___m

Task 1	Task time: ____s	❑ Successful	❑ Unsuccessful
Notes: …… …… …… ……			
Task 2	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……
Task 3	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……
Task 4	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……
Task 5	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……
Task 6	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……
Task 7	Task time: ____s	❑ Successful	❑ Unsuccessful
Preliminary problems discovered			Time problem occurred
…… …… …… …… ……			…… …… …… …… ……

Appendix C19: Payment Receipt

Incentive receipt and Acknowledgment Form

I hereby acknowledge receipt of £5 for my participation in a research study run by Mr. Obead Alhadreti.

Printed name: _____

Signature: _____

Date:_____

Appendix C20: Usability Problems Discovered

	Usability Problems	Identified by
1	The website does not support undo such as retrieving deleted records on E-shelf page.	CTA, HB
2	The 'Creation Date' function is limited from 1999 to 2013. The users cannot use this function for earlier dates	CTA, RTA, HB
3	The user expected the library catalogue to provide a short list of items recently searched on the catalogue page, but it was not there	RTA
4	The users cannot use the date sorting function tool in advanced search tool to specify publication dates other than listed	HB
5	The users cannot use the Boolean operators with more than two options (fields)	CTA, RTA, HB
6	The users expect an option on 'Catalogue' page to specify how many items to load per page	CTA
7	Find e-journal function returns some irrelevant results that did not met the users search terms	CTA, RTA, HB
8	Two main search tools on the homepage are confusing. The users thought the main search is for the library. However, it is for the website	CTA, RTA, HB
9	The users have to agree to the conditions each time they return to the 'Room Booking' page even if they had just done this. The website does not seem to save this.	CTA
10	Journal articles appear in the search results for "books and more" and vice versa, so "article search" seems redundant	CTA, RTA, HB
11	The sorting function for 'Books and more' returns irrelevant search results that did not met the users search terms	CTA, RTA, HB
12	The users feel that the help function is not sufficiently comprehensive	HB
13	There is no error message when the users do not enter search term in the website main search tool	HB
14	Some pages open in new tab, some open in same tab, confusing for the user	CTA
15	No "back to top" button after scrolling down a long way on the 'Additional Services' page	HB
16	The word 'Guest' on the 'Catalogue' page looks like clickable but it is not	HB
17	Too much scrolling on 'Additional Services' page	RTA
18	The site map is not easy to locate	CTA
19	The user feels that the font used in the 'Rooms Booking' page is too small	HB
20	Users failed to spot 'First' and 'Last' buttons for search results on 'Catalogue' page	CTA
21	The items do not show that they have been clicked on 'Catalogue page'	HB
22	No indication of the required or optional fields in the catalogue advanced search form	RTA, HB
23	The purpose of blue boxes is not clear on the 'Rooms Booking' page	HB
24	The button 'Action' is not easy to locate on 'Catalogue' page	CTA
25	Zooming images make them burlled	RTA
26	The font size of the link 'Advanced search' is too small	HB
27	There is no consistency in the font use in the left side bar on the catalogue page	CTA
28	On searching for the book through the main "search", the sidebar changes which is very confusing	CTA, RTA, HB
29	After clicking on a dropdown in the 'Browse Search' page, it remains open unless the user clicks on the same dropdown again	HB
30	There is no exist button from the error message in the 'Room Booking' page	HB
31	Users did not expect scrolling on the 'homepage' of the library	CTA, RTA, HB
32	There are two asterisks next to the first name filed on 'More Book' page which confused the users.	CTA, RTA
33	There is no consistency for the options displayed next to each item on the 'Catalogue' page	CTA, HB

34	The link 'Sign in to your library account' in the home page should be positioned at the top of the library 'Homepage'	CTA, HB
35	The search tool 'Search the library' should be positioned at the top	CTA, HB
36	The site did not arrange the results according to how relevant they were to the user's search terms	CTA, RTA, HB
37	No enough spaces between items on the 'E-Shelf' page	CTA, HB
38	Background and text colours are not appropriate	CTA, HB
39	Tooltips show just the names of the link, but no descriptions offered (e.g., LL request)	CTA, HB
40	The link 'Library Search' on the 'Homepage' is not clearly visible	RTA, HB
41	The site makes it hard to correct errors by positioning the cursor at the location where correction is not required	RTA, HB
42	The user did not know the meaning of the term 'ROAR'.	HB
43	The user did not know the meaning of the term 'Review and tag'	HB
44	The users did not understand the meaning of the term 'RSS' in the catalogue page	HB
45	The terms and conditions in the Booking a study rooms page are not clear	RTA
46	The user did not know the meaning of the term 'LibCal' on 'Rooms booking' page.	HB
47	The user did not know the meaning of the term 'E-shelf' on 'catalouge' page	RTA
48	The users are not sure from the instruction given whether or not they need to login in before booking a study room	CTA
49	The error message in 'Browse Search' page is not clear	HB
50	The user did not know the meaning of the term 'Periodical'	RTA
51	The user finds that the order of information is problematic in the 'Help and Support' page	CTA, HB
52	The information on the 'Study Support' page is not clearly structured	CTA, RTA
53	The user expected the citation option to be displayed with the item 'Details' section, but it was not there.	CTA
54	Users though that some information is repeated on different pages	CTA, RTA, HB
55	Too many images on the 'Study Support Page'	CTA
56	There is no direct link to go back to library's main page in the 'Rooms Booking' page	CTA, HB
57	Clicking on the library logo takes the user to the university home page instead of the homepage of the library	CTA, HB
58	No direct link to the help function in 'Rooms Booking' page	CTA, HB
59	There is no site map on each page	CTA, RTA, HB
60	The link 'Get it' is problematic because user though that they by clicking on this link they can view an electronic copy of item	CTA, RTA, HB
61	The link 'E-shelf' is problematic because many users failed to click on it to find information about their search history	CTA, RTA, HB
62	The link 'Action' is problematic because many users failed to click on it to find information about items citation	CTA, RTA, HB
63	The link 'Browse Search' was problematic because quite few users mistakenly thought that they could find information about their search history under this link.	CTA, RTA, HB
64	The link 'Go' in E-shelf page confused user	CTA, HB
65	The link 'My Account' was problematic because quite few users thought they can find information regarding their search history under this link	CTA
66	Users felt that the links 'My Account' and 'Sign In' are confusing because they seem to lead to functions that do the same thing.	CTA, RTA, HB
67	The labels of the links 'Subject Support' and 'Ask-A-librarian' confused the users because they were not sure which one to choose about subject support	CTA, RTA, HB
68	The labels of links 'Basket' and 'E-Shelf' confused the user because they look similar	CTA
69	The links 'Continue' and 'Submit Time Slots' on 'Rooms Booking' page confused the they are not sure which one to choose	CTA, RTA, HB

70	Clicking on the button 'Exist' on 'Rooms Booking' page take users to the library catalogue instead of returning them to the home page.	CTA
71	There is no direct link to the main search tool on 'Booking Rooms' page	RTA
72	The link containing the title of the items on the 'Catalogue' page is confusing because users think that by clicking the images will display full details on the item	CTA, RTA, HB
73	The images displayed next to each item after conducting search on 'Catalogue page' are confusing because users think clicking the images will display full details about the item, but they take users to almost empty pages	RTA
74	Users expect the link to FAQ to be with the main menu, but it was not there	HB
75	The labels of the links 'Subject Support' and 'Ask-A-librarian' confused the users because they were not sure which one to choose to find information about subject support	CTA, RTA, HB

Appendix C21: Appreciation Letter from the Administrator of the Website

University of East London

Dear Sir/Madame,

This is to confirm I have received the usability Problems report from Obead Alhadreti regarding our website.

Please be aware that during the last year UEL took the decision to split their website into an Internet – for external facing publicity and an Intranet for current students and staff. Most of the library and learning services website now sits within the Intranet.

Regards,

Victoria Voice
Electronic Services Librarian

Docklands Campus Library
University of East London
University Way
London
E16 2RD
v.voice@uel.ac.uk

Appendix C22: Normality Tests for the Experience with TA Test Questionnaire Data

	Shapiro-Wilk Test		
	CTA	**RTA**	**HB**
Working condition			
Slower than normal	.025	.054	.001
More focused	.008	.001	.007
Think-aloud experience			
Difficult	.023	.012	.001
Unnatural	.001	.004	.008
Unpleasant	.025	.001	.033
Tiring	.020	.007	.010
Time-consuming	.011	.051	.038
Evaluator presence			
Unnatural	.000	.005	.000
Disturbing	.000	.000	.000
Unpleasant	.000	.000	.000

Appendix C23: Normality Tests for Usability Problem Data

Individual problems	Shapiro-Wilk Test		
	CTA	RTA	HB
Number of individual problems	.417	.386	.037
Observed problems	.000	.000	.000
Verbalized problem	.009	.000	.014
Combination of both	.661	.031	.178
Critical problems	.000	.013	.000
Major problems	.422	.020	.002
Minor problems	.058	.003	.031
Enhancement problems	.000	.000	.000
Navigational problem	.002	.071	.115
Layout problems	.077	.007	.004
Content problems	.000	.000	.000
Functionality problems	.001	.000	.000

Appendix D: Materials from Study Two

Appendix D1: UEA Approval

Dear Obead,
The submission of your proposal has been considered by the UEA Computing Sciences Research Ethics Committee and I can confirm that your proposal has been approved.

Please could you ensure that any further amendments to either the protocol or documents submitted are notified to me, as Chair of CMP-REC, in advance and also that any adverse events which occur during your project are reported to the Committee.

The Committee would like to wish you good luck with your project

Best wishes,
Dan Smith
(Chair CMP-REC)

Dr D.J. Smith email: Dan.Smith@uea.ac.uk
School of Computing Sciences tel: +44 (0)1603 592608
University of East Anglia

Appendix D2: Email Sent to the Administrator of the Website

Dear Sir or Madam,

Thank you for your interest in my email.

My name is Obead Alhadreti, and I am a PhD student in human-computer interaction in the School of Computing Sciences at the University of East Anglia.

I am writing to you to seek your kind approval to use your University library's website as a test object in my usability study. The aim of this study is to evaluate the ease of use and user-friendliness of your library website. By assessing the usability of your website, I will better understand users' needs and expectations regarding academic library websites.

At the end of the study, I will provide you with a report that lists all usability problems found on your website.

I look forward to hearing from you at your earliest convenience.

Best regards,

Obead Alhadreti

Appendix D3: Website's Administrator Approval

Dear Obead Alhadreti,

Thank you for your e-mail and phone call. We would be quite happy for you to conduct a usability study on the Durham University library web site (and the Library catalogue and Discover systems if appropriate) providing you limit your study to those parts of the site which are accessible to the public. Issuing temporary access to the private parts of the site would be too tricky to arrange, but I understand that fits with what you had been intending anyway.

You asked about which services or sections of the site were the most used. I think it would be best to contact my colleague, Adam Walsh, about this. I am copying him in to this message. I expect he would be able to list the top twenty or so pages which should give you a good start.

You asked for permission to conduct the study in writing on headed paper, to satisfy your ethics processes. I am happy to do this if you will provide a postal address to which we can send the letter.

Best wishes,

--
Matthew Phillips

Appendix D4: Task List

Task ID	Task Description
T1	You have borrowed a laptop from the library of Durham University for 4 hours, but it turned out that you needed to use it for six hours instead. Using the website, please find the charge for late return. Can you find it?
T2	You want to find how many local studies the library catalogue has on the topic 'pollution'? Can you find them?
T3	You are taking a course on 'Web Technology'. Using the site, find the reading list for the course. Can you find it?
T4	You are a big fan of the author "Harriet Bulkeley" and want to know how many publications are written by your favourite author on the subject 'Climate change'. Can you find it?
T5	You want to book a room at the library to study for your coming exam. Using the website, find what the maximum time that you can book an individual room for. Can you find it?
T6	You are a first year PhD student in Law at the department of Law at Durham University and want to find all PhD thesis that have the key word "law" in the title in department of Law at Durham University. Can you find it?
T7	You want to find how many publications that have the keyword "usability" in their titles were published between 2010 and 2015. Can you find them?
T8	You are a part-time student who work off-campus for most of the time. You want to know what services the library offers for off-site users. Can you find them?
T9	You want to find how many publications the library catalogue has on the topic 'language', excluding the language 'English'.

Appendix D5: Consent Form

Consent Form

(Please read and sign this form)

University of East Anglia

Thank you for agreeing to participate in this study. The aim of this study is to evaluate a university library website. You will be encouraged to share your thoughts by thinking aloud (I will explain later what do I mean by this). During the study, it will be necessary for me to record a number of things using screen capture software, video and audio. However, this recorded data will be stored securely on a password-protected computer in accordance to the UEA's data protection policy. You can withdraw from this study at any time. Recording and notes taken will be destroyed as you require.

Please tick the box for things that you agree with, and sign below if you are happy to give your consent for the study to go ahead.

1. Your monitor and voice will be recorded using screen capture. [ ]
2. The results of the analysis of this evaluation may be published, but all the data recorded will be anonymous. [ ]

* If you would like to access to any reports or publications that directly result from your involvement in this study, please tick the box. [ ]

Participant Name	*Signature of Participant*	*Date*
_____	_____	___ / ___ /201

Contact details:

Researcher:	Obead Alhadreti	Supervisor: Dr. Pam Mayhew
Email address:	O.Alhadreti@uea.ac.uk	P.Mayhew@uea.ac.uk
Contact number:	07923206416	01603593334

Appendix D6: Procedure Sheet

Hi and thank you for coming today. As you probably know my name is 'Obead'.

1. Please review and sign the informed consent form which will provide you with an overview of the study.

2. Please take a moment to familiarize yourself with the laptop and Internet browser.

3. Thinking Aloud: 'In this study, I am interested in what you say to yourself as you perform some tasks that I give you. In order to do this I will ask you to think aloud as you work on the tasks. What I mean by think-aloud is that I want you to say out loud everything that you say to yourself silently. Just act as if you are alone in the room speaking to yourself. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?'

4. Let's take a moment to practice thinking aloud. Please think aloud as you look up for the word 'chant' in the online dictionary opened in the browser. Please note that this is not the website under evaluation.
   Do you have any questions about the thinking aloud process you've just practiced?

5. Please read carefully task instructions on the screen.

6. If you don't have any questions, please start performing the first task written on the note card, placed on the table on your right hand, using the website. Please remember to think aloud while you are solving the tasks from beginning till the end of the task.

7. Now you have finished the tasks, please answer the online post-experiment questionnaires.

Thank you very much for your time and input. Here is your reward. Have a great day!

Appendix D7: Intervention List

Intervention type	Intervention Trigger
Reminder	Participants fall silent more than 15 seconds, the evaluator reminds them to keep talking.
Clarification	When participants solve the task with unclear goals or actions; when participants make vague comments, the evaluator asks for a clarification
Ask Explanation	Participants express difficulties, feelings, likes, dislikes etc., without giving an explanation, the evaluator asks for an explanation
Interjection Exploration	Participants make an interjection but no further comments
Seek Opinion	Participants give an evaluation summary of information or outcome of their actions in the areas which may have a potential problem, the evaluator asks about the user experience and ease of task in general
Ask Suggestion	Participants verbalise difficulties or negative feelings, the evaluator asks for suggestions
User Expectation	Participants indicate something does not meet their expectations, the evaluator enquires about their expectations
Task Continuation	Participants think the task is finished; Participants are too chatty; Participants misunderstood the task; Participants give up too easily. Participants become frustrated,

Adopted from (Zhao and McDonald, 2010; Naveedh, 2015).

Appendix D8: Observation Sheet

Usability Test Observation Sheet

Participant #: ____	**TA method:** ____	**Date:** / /201
Session starts at: ___ h ___m		**ends at:** ___ h ___m

Task 1 Notes: …… …… …… ……
Task 2 Notes: …… …… …… ……
Task 3 Notes: …… …… …… ……
Task 4 Notes: …… …… …… ……
Task 5 Notes: …… …… …… ……
Task 6 Notes: …… …… …… ……
Task 7 Notes: …… …… …… ……
Task 8 Notes: …… …… …… ……
Task 9 Notes: …… …… …… ……

Appendix D9: Usability Problems Discovered

	Usability Problems	Identified by
1	On the catalogue page, when a search is performed with an empty search box, no error message is displayed.	CTA, SC, AI
2	On the 'Advanced Search' page, users' language choices are restricted to either all languages or a single specified language. Users cannot select and exclude multiple languages.	CTA, SC, AI
3	When a word is misspelled (e.g. 'polluton'), no error message is displayed.	AI
4	If no results are found there is no feedback or notification stating this, which makes users doubtful about the efficacy of the search process.	CTA, SC, AI
5	There is no validation of data entered into the advanced search (particularly the 'from' and 'to' date fields).	AI
6	There are two search boxes on the home page, which seems to be confusing	CTA, SC, AI
7	The search engine does not provide automatic spellchecking, which some users find frustrating.	SC
8	No messages displayed in some pages when there are observable delays. Users want ongoing updates.	AI
9	There is no validation of data entry on the 'More Books' page, which impacts on the accuracy of the search process.	CTA, SC, AI
10	There is no option to sort items by publisher, which makes the search process longer and more laborious than necessary.	CTA, SC, AI
11	The filter on the 'Modify Search' page doesn't give relevant results. When users opt to sort items by relevance, the results displayed are sorted by date.	CTA
12	On the 'Basic Search' page, there is no option to sort results by author, which makes the search process longer than necessary.	CTA, SC, AI
13	The users do not get a feedback after clicking on "Book Cart" to informing him what to do next, which seems to be confusing.	SC
14	On the 'Modify Search' page, there is no option to sort results by author.	SC, AI
15	The user does not receive any feedback or confirmation after submitting a search request on the 'More Books' page, which makes the user uncertain.	AI
16	The site does not allow the users to recover from errors by showing confirmation message such as after deleting an item in the "Book Cart" page.	AI
17	The number of results per page cannot be configured by the user, which make the search process unpleasant for the user.	SC, AI
18	There is no option to clear the form in "Modify Search" page. The user had to clear each filed individually.	SC
19	On "Search reading lists" page, there is no information regarding when the item was published. The user thought it is necessary.	CTA, SC
20	The term 'limited' on the Reading Lists' page is not adequately explained	CTA
21	Users do not understand the meaning of the term 'ShelfMark'. This unfamiliar word is not explained properly.	CTA, SC
22	Users do not know the meaning of the term 'Repositories'.	AI
23	The 'Book Cart' page does not display the dates that items were added to the cart.	CTA
24	On the 'Borrowing Laptops' page, there is no information regarding the loan period, which is necessary when borrowing an item from the library.	CTA
25	On the 'Borrowing Laptops' page, there is no information regarding when the page was last updated.	AI
26	Users are uncertain whether to use an author's first or second name when searching.	CTA
27	On the 'Modify Search' page, when an incorrect date is entered, the error message does not clearly state the problem.	CTA, AI
28	The 'GO' button on the 'Simple Search' page is not properly worded. It should be 'Search'.	SC, AI
29	Some information is repeated across different pages.	CTA

30	No explanation of the difference between subject support and a librarian, which There is no explanation of the difference between 'subject support staff' and a librarian, which users found confusing.	SC
31	The names of submenus are too long.	CTA, AI
32	The 'Advanced Search' page cannot be accessed directly from the home page. The participants thought it should be more easy to reach.	CTA
33	Bookable study rooms are listed under the 'Services and Site' section on the left navigation bar. Participants thought it would be better if they were under the 'Reserve' section.	CTA, SC, AI
34	The results page contains two buttons, one reading 'Start Over', and another reading 'Another Search'. Participants found this confusing and were unsure of which to choose.	CTA, SC, AI
35	The wording of the 'Modify Search' button is confusing. It should be changed to 'Advanced Search'.	CTA, SC, AI
36	There is no site map on every page, which can make users feel lost.	CTA, SC, AI
37	A link to the library account is not clearly visible on every page.	CTA, SC, AI
38	There is no direct link to the FAQ on the home page	CTA
39	Hypertext links that invoke actions are not clearly distinguished from hypertext links that load another page.	SC, AI
40	Some links take the user back to the same page (circular links).	CTA
41	Clicking on the logo takes the user to the university home page instead of the library home page.	CTA, SC, AI
42	No direct link to the library main page in the "Reading List" page	SC
43	The users found fault with the order of the links on the 'Quick Link' section. The likely task priorities of actual users do not seem to have been considered.	AI
44	The 'Book an individual or group study room' link is not clearly visible.	CTA, SC, AI
45	The 'New Search' and 'Simple Search' buttons are too similar and confuse the user.	AI
46	The 'Repositories' link is not clearly positioned.	CTA, SC, AI
47	There is no link to the 'Help' page on the library catalogue page	AI
48	The 'Essential Info' and 'Information for Students' links are confusing. Users were unsure which one to choose in order to access information about services for part-time students	CTA, SC, AI
49	On the 'e-Theses' page, there is no direct link back to the main page.	CTA, SC, AI
50	Some text is clickable but is not visibly clickable, so users might not see it	CTA, SC, AI
51	Some pages lack navigational feedback showing users where they are in the site	CTA, SC, AI
52	The 'Repositories' link is problematic—many users did not expect to be able to search for e-theses in this section.	CTA, SC, AI
53	The main menu and the sub menu are very close together on some pages, making navigation difficult.	CTA, SC, AI
54	There is no clickable indication of the current page in the secondary navigation.	CTA, AI
55	The home page has too many menus and sub-menus, making it difficult to scan	CTA, SC, AI
56	The right-hand side navigation menus are inconsistent: every webpage has different tabs.	CTA, SC, AI
57	Low colour contrast on the results page, making it difficult to read.	CTA, SC, AI
58	Too many results per page, leading to excessive scrolling.	CTA, SC, AI
59	My reading list webpage is inconsistent with others webpages in the library website, for example the header and footer disappeared as well as the main menu	SC
60	On the search page there are two text boxes, one to 'Search by Course' and the other to 'Search by Lecturer'. However, on the results page there is just one text box, and a secondary dropdown list to filter the search by either course or lecturer.	SC
61	There are too many dropdown lists on the 'Modify Search' page.	CTA, SC, AI
62	On the 'Modify Search' page, the 'Submit' button should be positioned under the search form rather than next to it	CTA, SC, AI
63	On the 'Study Room' page, important details (e.g. the maximum length of time that students can book rooms) are not highlighted sufficiently.	CTA, SC, AI

64	There are too many text boxes, dropdown lists, and checkboxes on the 'Search e-Theses' page.	CTA, SC, AI
65	The user interface does not look very attractive.	CTA
66	Link colours should be standardised.	SC
67	On data entry forms, the cursor is not placed where the input is needed	SC
68	The 'New Search' button is too small.	CTA
69	The 'Search Results' page does not clearly state the number of results retrieved.	CTA, SC, AI
70	Some pages have 'scroll stoppers' (headings or page elements that create the illusion that users have reached the top or bottom of a page when they have not).	CTA
71	On the catalogue page, the 'First Page' and 'Last Page' buttons should be positioned at the top.	SC
72	The link to the library's home page is not located in the same place on every page.	CTA
73	The information on the 'Special Collections' page is not clearly structured.	SC
74	There is too much information on some pages.	AI
75	The image on "your library account" page is not clearly visible	AI
76	The image on the 'Your Library Account' page becomes blurred when zoomed in.	CTA
77	On the 'Basic Search' page, the options in the dropdown menu are not ordered logically or alphabetically.	CTA, SC, AI
78	Some of the information about services for part-time student services is provided as PDF documents rather than webpages, though not all browsers support in-line PDF viewing.	CTA
79	On the 'Basic Search' page, the dropdown list is too long.	CTA
80	The 'Advanced Search' form gives no indication of which fields are required and which are optional.	CTA, SC, AI
81	On the 'Catalogue' page, the 'Library Homepage 'link should be positioned to the left and the 'Other Library' link to the right.	SC
82	Menu and submenu labels do not offer any descriptions.	AI
83	On the home page, there is no clearly visible means for users to log in to their library account. A login link should be clearly positioned.	SC, AI
84	Some pages (e.g. 'Special Collections') require excessive scrolling.	CTA, AI
85	Text entry fields do not indicate the amount of data that needs to be entered	AI
86	Pages with excessive scrolling to not provide a 'Back to Top' link.	CTA, SC
87	On the catalogue pages, the results and tables should be better-presented and more structured.	CTA, AI
88	There is no consideration for accessibility features such as using alternative texts for image to displayed when system response is slow	AI
89	There is no consistency in the shape and the size of buttons, e.g. on the 'Modify Search' page	CTA, SC
90	Home page needs scrolling—it is difficult to see it in a single glance.	CTA, SC, AI
91	Some tick-box labels are partially overlapping	AI
92	No enough spaces between search options	SC, AI
93	There is no 'Add to e-Shelf' button on individual item pages.	CTA, SC
94	The opening hours of the advice centre is not positioned clearly	AI
95	In general, the layout does not help users to focus attention on what to do next.	SC, AI
96	On some pages, there is horizontal scrolling, which some users did not like.	CTA, SC
97	The fields on the 'Advanced Search' page are not appropriate to the size of the input.	AI
98	The page titles are duplicated on some pages.	AI

Appendix D10: Normality Tests for the Experience with TA Test Questionnaire Data

	Shapiro-Wilk Test		
	CTA	**SC**	**AI**
Working condition			
Slower than normal	.024	.007	001
More focused	.000	.000	.045
Think-aloud experience			
Difficult	.000	.000	.006
Unnatural	.006	.007	.016
Unpleasant	.033	.000	.029
Tiring	.001	.000	.001
Time-consuming	.000	.002	.016
Evaluator presence			
Unnatural	.000	.000	.000
Disturbing	.000	.000	.011
Unpleasant	.000	.000	.000

Appendix D11: Normality Tests for Usability Problem Data

Individual problems	Shapiro-Wilk Test		
	CTA	SC	AI
Number of individual problems	.043	.341	.783
Observed problems	.017	.015	.281
Verbalized problem	.075	.088	.009
Combination of both	.045	.381	.255
Critical problems	.011	.002	.002
Major problems	056	.299	.128
Minor problems	.001	.068	.018
Enhancement problems	.000	.000	.000
Navigational problem	.015	.015	.003
Layout problems	.009	.040	.281
Content problems	.000	.000	.000
Functionality problems	.009	.017	.200

Appendix E: Materials from Study Three

Appendix E1: UEA Approval

Dear Obead,
The submission of your proposal has been considered by the UEA Computing Sciences Research Ethics Committee and I can confirm that your proposal has been approved.

Please could you ensure that any further amendments to either the protocol or documents submitted are notified to me, as Chair of CMP-REC, in advance and also that any adverse events which occur during your project are reported to the Committee.

The Committee would like to wish you good luck with your project

Best wishes,
Dan Smith
(Chair CMP-REC)

Dr D.J. Smith email: Dan.Smith@uea.ac.uk
School of Computing Sciences tel: +44 (0)1603 592608
University of East Anglia

Appendix E2: Co-participation Procedure Sheet

Hi and thank you for coming today. As you probably know my name is 'Obead'.

1. Please now review and sign the informed consent forms which will provide you with an overview of the study.

2. Please take a moment to familiarize yourself with the laptop and Internet browser.

3. "In this study, I am interested how you solve some tasks that I give you. Even though only one of you can actually control the mouse, you have to perform the tasks as a team by consulting each other and making joint decisions. I also want you to state aloud what you are doing. If you are silent for any length of time I will remind you to keep talking aloud. Do you understand what I want you to do?"

4. Let's take a moment to practice this. Please work together as you look up for the word 'carol' in the online dictionary opened in the browser. Please note that this is not the website under evaluation. Do you have any questions about the process you've just practiced?

5. Please read carefully task instructions on the next page.

6. If you don't have any questions, please start performing the tasks.

7. Now you have finished the tasks, please answer the online post-experiment questionnaires.

Thank you very much for your time and input. Here are your rewards. Have a great day!

Appendix E3: Usability Problems Discovered

	Usability Problems	Identified by
1	On the catalogue page, when a search is performed with an empty search box, no error message is displayed.	CTA & CP
2	On the ‘Advanced Search’ page, users’ language choices are restricted to either all languages or a single specified language. Users cannot select and exclude multiple languages.	CTA & CP
3	When a word is misspelled (e.g. ‘polluton'), no error message is displayed.	CP
4	If no results are found there is no feedback or notification stating this, which makes users doubtful about the efficacy of the search process.	CTA & CP
5	There is no validation of data entered into the advanced search (particularly the 'from' and 'to' date fields).	CP
6	There are two search boxes on the home page, which seems to be confusing	CTA & CP
7	The search engine does not provide automatic spellchecking, which some users find frustrating.	CP
8	There is no validation of data entry on the 'More Books' page, which impacts on the accuracy of the search process.	CTA & CP
9	There is no option to sort items by publisher, which makes the search process longer and more laborious than necessary.	CTA &CP
10	The filter on the 'Modify Search' page doesn't give relevant results. When users opt to sort items by relevance, the results displayed are sorted by date.	CTA
11	On the ‘Basic Search’ page, there is no option to sort results by author, which makes the search process longer than necessary.	CTA &CP
12	On the 'Modify Search' page, there is no option to sort results by author.	CP
13	The user does not receive any feedback or confirmation after submitting a search request on the 'More Books' page, which makes the user uncertain.	CP
14	Users can not use parts of words, e.g. ‘pollutio’ or ‘ollution’.	CP
15	There is no indication of how many copies of each item are available.	CP
16	On the 'Search Reading Lists' page, there is no information regarding when items were published.	CTA & CP
17	The term 'limited' on the Reading Lists' page is not adequately explained	CTA & CP
18	Users do not understand the meaning of the term 'ShelfMark'. This unfamiliar word is not explained properly.	CTA & CP
19	Users do not know the meaning of the term ‘'Repositories'.	CP
20	The ‘Book Cart’ page does not display the dates that items were added to the cart.	CTA
21	On the 'Borrowing Laptops' page, there is no information regarding the loan period, which is necessary when borrowing an item from the library.	CTA
22	On the 'Borrowing Laptops' page, there is no information regarding when the page was last updated.	CP

23	Users do not understand the meaning of the term 'SCONUL'	CP
24	Users are uncertain whether to use an author's first or second name when searching.	CTA
25	On the 'Modify Search' page, when an incorrect date is entered, the error message does not clearly state the problem.	CTA & CP
26	The 'GO' button on the 'Simple Search' page is not properly worded. It should be 'Search'.	CP
27	Some information is repeated across different pages.	CTA
28	Users do not understand the meaning of the term 'COPAC'	CP
29	There is no explanation of the difference between 'subject support staff' and a librarian, which users found confusing.	CP
30	The names of submenus are too long.	CTA & CP
31	There is no indication of when pages were last updated	CP
32	The 'Advanced Search' page cannot be accessed directly from the home page. The participants thought it should be more easy to reach.	CTA
33	Bookable study rooms are listed under the 'Services and Site' section on the left navigation bar. Participants thought it would be better if they were under the 'Reserve' section.	CTA & CP
34	The results page contains two buttons, one reading 'Start Over', and another reading 'Another Search'. Participants found this confusing and were unsure of which to choose.	CTA & CP
35	The wording of the 'Modify Search' button is confusing. It should be changed to 'Advanced Search'.	CTA & CP
36	There is no site map on every page, which can make users feel lost.	CTA & CP
37	A link to the library account is not clearly visible on every page.	CTA & CP
38	There is no direct link to the FAQ on the home page	CTA
39	Hypertext links that invoke actions are not clearly distinguished from hypertext links that load another page.	CP
40	Some links take the user back to the same page (circular links).	CTA
41	Clicking on the logo takes the user to the university home page instead of the library home page.	CTA & CP
42	The site has orphan (dead-end) pages.	CP
43	The users found fault with the order of the links on the 'Quick Link' section. The likely task priorities of actual users do not seem to have been considered.	CP
44	The 'Book an individual or group study room' link is not clearly visible.	CTA & CP
45	The 'New Search' and 'Simple Search' buttons are too similar and confuse the user.	CP
46	The 'Repositories' link is not clearly positioned.	CTA & CP
47	There is no link to the 'Help' page on the library catalogue page	CP

48	The ‘Essential Info’ and ‘Information for Students’ links are confusing. Users were unsure which one to choose in order to access information about services for part-time students	CTA & CP
49	On the ‘e-Theses’ page, there is no direct link back to the main page.	CTA & CP
50	Some text is clickable but is not visibly clickable, so users might not see it	CTA & CP
51	Some pages lack navigational feedback showing users where they are in the site	CTA & CP
52	The 'Repositories' link is problematic—many users did not expect to be able to search for e-theses in this section.	CTA & CP
53	The main menu and the sub menu are very close together on some pages, making navigation difficult.	CTA & CP
54	There is no clickable indication of the current page in the secondary navigation.	CTA & CP
55	The home page has too many menus and sub-menus, making it difficult to scan	CTA & CP
56	The right-hand side navigation menus are inconsistent: every webpage has different tabs.	CTA & CP
57	Low colour contrast on the results page, making it difficult to read.	CTA & CP
58	Too many results per page, leading to excessive scrolling.	CTA & CP
59	On the search page there are two text boxes, one to ‘Search by Course’ and the other to ‘Search by Lecturer’. However, on the results page there is just one text box, and a secondary dropdown list to filter the search by either course or lecturer.	CP
60	There are too many dropdown lists on the 'Modify Search' page.	CTA & CP
61	On the ‘Modify Search’ page, the ’Submit' button should be positioned under the search form rather than next to it	CTA & CP
62	On the 'Study Room' page, important details (e.g. the maximum length of time that students can book rooms) are not highlighted sufficiently.	CTA & CP
63	There are too many text boxes, dropdown lists, and checkboxes on the 'Search e-Theses’ page.	CTA & CP
64	The user interface does not look very attractive.	CTA
65	Link colours should be standardised.	CP
66	On data entry forms, the cursor is not placed where the input is needed.	CP
67	The 'New Search' button is too small.	CTA
68	The ‘Search Results’ page does not clearly state the number of results retrieved.	CTA & CP
69	The site uses italicised text, which is not preferred by users.	CP
70	Some pages have 'scroll stoppers' (headings or page elements that create the illusion that users have reached the top or bottom of a page when they have not).	CTA
71	On the catalogue page, the 'First Page' and ‘Last Page’ buttons should be positioned at the top.	CP
72	The image on the catalogue page looks as if it is clickable, but it is not.	CP

73	The link to the library's home page is not located in the same place on every page.	CTA
74	The information on the 'Special Collections' page is not clearly structured.	CP
75	There is too much information on some pages.	CP
76	The image on the 'Your Library Account' page becomes blurred when zoomed in.	CTA & CP
77	On the 'Basic Search' page, the options in the dropdown menu are not ordered logically or alphabetically.	CTA & CP
78	On the 'University Library ConneXions' page, the colour of the options in the navigation bar make them hard to read.	CP
79	Some of the information about services for part-time student services is provided as PDF documents rather than webpages, though not all browsers support in-line PDF viewing.	CTA
80	On the 'Basic Search' page, the dropdown list is too long.	CTA & CP
81	The information in the section 'About the University Library and Heritage Collections' is not clearly structured.	CP
82	The 'Advanced Search' form gives no indication of which fields are required and which are optional.	CTA & CP
83	On the 'Catalogue' page, the 'Library Homepage 'link should be positioned to the left and the 'Other Library' link to the right.	CP
84	Menu and submenu labels do not offer any descriptions.	CP
85	On the home page, there is no clearly visible means for users to log in to their library account. A login link should be clearly positioned.	CP
86	Some pages (e.g. 'Special Collections') require excessive scrolling.	CTA & CP
87	Pages with excessive scrolling to not provide a 'Back to Top' link.	CTA & CP
88	On the catalogue pages, the results and tables should be better-presented and more structured.	CTA & CP
89	There is no consistency in the shape and the size of buttons, e.g. on the 'Modify Search' page	CTA & CP
90	Home page needs scrolling—it is difficult to see it in a single glance.	CTA & CP
91	There is not enough space between search options.	CP
92	There is no 'Add to e-Shelf' button on individual item pages.	CTA & CP
93	In general, the layout does not help users to focus attention on what to do next.	CP
94	On some pages, there is horizontal scrolling, which some users did not like.	CTA & CP
95	The fields on the 'Advanced Search' page are not appropriate to the size of the input.	CP
96	The page titles are duplicated on some pages.	CP

Appendix E4: Normality Tests for the Experience with TA Test Questionnaire Data

	Shapiro-Wilk Test	
	CTA	**CP**
Working condition		
Slower than normal	.024	.072
More focused	.000	.001
Think-aloud experience		
Difficult	.000	.001
Unnatural	.006	.001
Unpleasant	.033	.000
Tiring	.001	.001
Time-consuming	.000	.000
Evaluator presence		
Unnatural	.000	.001
Disturbing	.000	.000
Unpleasant	.000	.000

Appendix E5: Normality Tests for Usability Problem Data

Individual problems	Shapiro-Wilk Test	
	CTA	CP
Number of individual problems	.043	.378
Observed problems	.017	.070
Verbalized problem	.075	.002
Combination of both	.045	.839
Critical problems	.011	.001
Major problems	056	.166
Minor problems	.001	.172
Enhancement problems	.000	.017
Navigational problem	.015	.015
Layout problems	.009	.357
Content problems	.000	.010
Functionality problems	.009	.221

Appendix F: Research Publications/Presentations/Activities List

During the period of this thesis, and in an effort to connect with a wide number of researchers in this field and gain their feedback regarding this area of research, the researcher accomplished the following:

- Published a number of papers,
- Delivered posters,
- Attended doctoral consortiums,
- Created a blog site containing a wealth of links that are very useful for researchers in this field. (address: http://tautm.wordpress.com/),
- Co-supervised six Master's projects.

Published Papers:

1. Alhadreti, O., Al Roobaea, R., Wnuk, K., Mayhew, P. J. (2014). The impact of usability of online library catalogues on user performance. In: IEEE, *International conference on information science and applications*. Seoul, Republic of Korea, 6-9 May 2014.
2. Alshammari, T., Alhadreti, O., & Mayhew, P. J. (2015). When to Ask Participants to Think Aloud: A Comparative Study of Concurrent and Retrospective Think-Aloud Methods. *International Journal of Human Computer Interaction (IJHCI)*, *6*(3), 48.
3. Alnashri, A., Alhadreti, O., Mayhew, P. J. (2016). The Influence of Participant Personality in Usability Tests. *International Journal of Human Computer Interaction (IJHCI)*, 7 (1), 1-22.
4. Alqahtani, M. A., Alhadreti, O., AlRoobaea, R. S., & Mayhew, P. J. (2015). Investigation into the impact of the usability factor on the acceptance of mobile transactions: Empirical study in Saudi Arabia. *Int. J. Hum. Computer. Interact*, *6*, 1-35.

Papers under Review:

5. Alhadreti, O., Mayhew, P. J., 2016. To Intervene or Not to Intervene: An Investigation of Three Think-Aloud Protocols in Usability Testing. *Journal of Usability Studies*.

6. Elbabour, F., Alhadreti, O., Mayhew, P. J., 2016. Eye Tracking in Retrospective Think Aloud Usability Testing: is there Added Value? *Journal of Usability Studies*.

Posters:

- "The Effect of Thinking Aloud on Usability Testing", at the Centre for Internationalisation and Usability, University of West London. January, 2014.
- "An Investigation of Think-aloud Methods in Usability Testing", *The 30th British Human Computer Interaction Conference*. Bournemouth University. July, 2016.

Doctoral Consortium Attended:

- British Computer Society (May, 2013), London.