# AN INVESTIGATION OF THE IMPACT OF TASK-TYPES ON THE REACTIVITY OF THE CONCURRENT THINK-ALOUD IN USABILITY TESTING

**OBRUCHE ORUGBO EMUEAKPORAVWA**

A thesis submitted in partial fulfilment of the
requirements of the University of Sunderland
for the degree of Doctor of Philosophy

June 2023

## ABSTRACT

The Concurrent think-aloud (CTA) is primarily used to understand users' task based cognitive processes. However, is not without limitations. CTA procedures varies widely among practitioners. Also, it has been known to cause reactivity: an artificial change in task performance. This is problematic because it may alter the accuracy of task performance. Also, research on reactivity within usability testing have shown mixed findings. Thus, conclusions cannot be drawn to attest to whether reactivity occurs due to varying administration procedures and therefore we must now consider its relationship to other test-based factors. This research will be the first to systematically investigate the impact of task-type on reactivity of the CTA and the first to systematically investigates practitioners working habit in terms of their views on reactivity when using CTA in practice. Three studies were conducted, the first study investigates the Impact of task-types on the Reactivity of CTA and uses a mixed design. The results suggest that, thinking aloud during usability testing does not cause reactivity, and task type does not impact concurrent think-aloud. However, sensemaking tasks increase mental demand. The second study investigates the impact of task-type on two different think-aloud protocols and uses a mixed design. The result indicates that, the classic think-aloud method led to more successful task completion and no reactivity, while the explicit instruction produced fewer successful task completions and a higher mental workload. The explicit instruction produced less verbalisation, resulting in fewer relevant explanatory utterances, contradicting expectations. The third study uses an interview method to explores practitioners' experiences, views on reactivity and challenges when using the think-aloud method within usability testing. These studies demonstrates unequivocally that CTA should not be abandoned in usability studies as it provided valuable think-aloud data and helped identify usability issues. Additionally, practitioners should not replace the traditional think-aloud approach with explicit instruction, as explicit instruction had a greater influence on participants' behaviour. Ericsson and Simon's recommendations should be used for concurrent data collection, as it ensures data validity and generates the same type of data as explicit instruction while reducing reactivity.

**ACKNOWLEDGEMENTS**

My heartfelt gratitude goes to my ex-supervisor and mentor, Sharon McDonalds, through her understanding of the subject, her insistence on clarity, her elaboration of the basic ideas, she has always been available to me for guidance and support, her meeting frequency, proof reading, feedback and how she tracks my progress have been very helpful in shaping my research. Without her, I would not have been able to complete what I did for my research project. She will always be an inspiration to me and to my future career.

I offer my sincere thanks to my supervisor, Kenneth McGarry, for his immerse contributions, invaluable feedback on statistics, proofreading and helpful meetings. I thank you for your support and guidance.

With great love and deepest gratitude to my wife, Susan, my daughters, Leona and Lucia. With unhesitating faith and support they encouraged me to pursue my lifelong dream to achieving my PhD, amidst all the demands of family life. Doing this research has indeed been disruptive to them in terms of time and absence from home, and I am forever grateful for their love and support. Susan, in particular, is the smartest and most patient person I have known. Whenever I went home perplexed with ideas in half-baked condition, I always ended up with clarified ideas through my conversations with her. She is a great colleague, supporter and friend. I dedicate this research to her and our children.

To my dad blessed memory. My mum, Mary and my sister, Jovita, I can thank you enough for your patience and understanding, words can explain how much I miss you both. To my uncle, Hycenth and Cyril who has inspired and led me to reach the goals that I continue to set for myself. To Peter who has helped me in numerous ways. I thank you all for your support and kind words.

The path I journeyed has afforded me the blessing of many great friendships, which grew out of the unexpected twists and turns along the way, Ibrahim and Sarah, Julius. They remind me there is yet so much love, kindness and compassion in this ever-chaotic world we share. It would be impossible to name everyone here but for everyone that supported me through this

journey, you are truly appreciated. Lastly, I want to say thank every member of staff especially the GRS team for their unwavering tolerance without any hint of complaint.

**Table of Contents**

**APPENDICES**

**LIST OF FIGURES**

**LIST OF TABLES**

# CHAPTER ONE

## 1.0 INTRODUCTION

### 1.1 HCI & Usability

The last two decades of the 21st century have certainly seen usability became one of the most representative terms of the human-computer interaction (HCI) field, a vital subject to both researchers and practitioners. Indeed, the term became so popular that it transcended HCI and has come to denote a quality of any "human-made object". Perhaps one of the reasons for the broad appeal of the term usability is its prevalent interpretation as the desired interactive attribute of ease-of-use or "user-friendliness." Although, usability both as a practice and as an emerging field in HCI, has had its share of controversies which has been inherited from its early roots in experimental psychology, measurement and statistic, others have emerged as a result of its advancement and extension into the user-centred design and user experience (Lewis, 2014).

Think aloud was introduced in systems development by Clayton H. Lewis at IBM in 1982, as one of many inventions from US and European cognitive psychology imported into the then-emerging field of HCI (Lewis, 1982). It involves a small number of users' that think out loud (expressing their thoughts as soon as they occur) while carrying out tasks on a tested product with the presence of a usability test facilitator who observes, listen to users' verbalisation and capturing users' screen activities, recording their utterances and other metrics on a specialised computer. This process is referred to as usability testing. The primary aim of think-aloud is to gain insight into users' task solving strategies and behaviours by obtaining a report on the user's experience when interacting with a design to identify usability problems and possible suggestions for improvement (Olmsted-Hawala et al., 2010; McDonald, Zhao and Edwards, 2013a). However, its is not without limitations, one of which is reactivity.

## 1.2 Reactivity

Reactivity refers to an artificial change in task performance which makes a usability test no longer a representation of real-world use. This can be problematic because it may alter the accuracy of task performance, making it difficult to draw accurate conclusions about the usability of the product. The concept of reactivity has been the subject of much debate in usability testing, with researchers offering conflicting opinions on its existence and significance. The variability in CTA procedures among practitioners has contributed to the mixed findings regarding reactivity in usability testing.

Empirical demonstrations of reactivity within usability testing have shown mixed findings. Studies with evidence of reactivity Van den Haak and de Jong, (2003); Eger et al., (2007) and Bowers and Snyder, (1990). Studies without evidence of reactivity Olmsted-Hawala et al., (2010); McDonald and Petrie, (2013) and McDonald, Zhao and Edwards, (2015).

The subjective nature of usability testing methods, and the lack of standardisation in CTA procedures, has made it difficult to draw firm conclusions about the existence and impact of reactivity. Therefore, there is a need to investigate the impact of task-type on reactivity of the CTA and the practitioners' working habit in terms of their views on reactivity when using CTA in practice.

This research is important because it has the potential to inform the use of CTA in usability testing, which is critical for the development of software and digital products. If reactivity can be better understood and controlled, the accuracy of data collected in usability testing can be improved, leading to better product design and development. Additionally, this research has implications for practitioners who use the think-aloud protocol in their work. The findings of this research can inform practitioners' approach to CTA and provide guidelines for its use in usability testing.

Studies have compared both classic and relaxed think aloud or explicit instructions, however, none has compared CTA; Explicit instruction and Silent working with Fact, and assessment tasks to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity.

Given the current situation, it is important for usability practitioners to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity. As a result, it seems that exploring the concurrent think-aloud and reactivity still need major research contributions to further establish what might influence reactivity, and this thesis aims to explore other factors towards making such contribution.

### 1.3 Research Aims

The aim of this PhD research is to bridge the gaps identified above on previous research of Concurrent think-aloud and reactivity by investigating the impact of task type on reactivity in CTA and practitioners' views on reactivity in usability testing. Specifically, the study aims to determine if varying administration procedures lead to reactivity and if reactivity is related to other test-based factors. Additionally, the study will explore the impact of task type on two different think-aloud protocols and how UX practitioners use the CTA method in industry. The study will also examine practitioners' views on reactivity and the challenges they face when using the think-aloud method in usability testing.

To achieve these objectives, the research will involve (a) a critical review of related literature, (b) methodological exploration based on past studies, (c) study design and piloting, (d) three studies, (e) analysis and discussion, (f) evaluation and conclusion, and (g) reflections and future work. These objectives are discussed in detail in Section 1.8 of this thesis.

## 1.4 Research Questions

To gain a more substantiated knowledge the following research questions were formulated for this study:

**Research question 1(RQ1):** What is the impact of task type on reactivity in the concurrent think-aloud?

i. Does the act of thinking aloud under classic administration procedures cause reactivity within usability testing?
ii. What is the impact of different types of tasks on the reactivity of CTA?
iii. What is the relationship between task type and CTA administration procedures?

**Research question 2(RQ2):** The second study will answer the following research questions:

i. What is the impact of task performance on the use of fact and assessment task with the classic think-aloud, explicit instruction or silent within usability testing?
ii. Does explicit instruction lead to high mental workload over classic think-aloud and Silent?
iii. Does explicit instruction lead to an increase in relevant explanatory utterances in terms of user experience and expectations?

**Research question 3(RQ3): What are the practices and challenges of using the think-aloud protocol in the industry?**

(i) How do UX practitioners use the think-aloud method within usability testing?

(ii) What is the nature of tasks practitioners uses?

(iii) What are practitioners' views on reactivity?

Summarises the research questions addressed in each empirical study of the PhD research.

| Empirical Studies | Research questions |
|---|---|
| Study One: The Impact of Task-types on the Reactivity of the Classic Think-Aloud | RQ1, |
| Study Two: The Impact of Task-Type on Two Different Think-aloud | RQ1, RQ2 |
| Study Three: Practitioners Use of Think-Aloud: Practises and Challenges | RQ3, RQ1, RQ2ii |

*Table 1: Empirical studies and research questions*

### 1.5 Impact of the COVID-19 Restrictions on My PhD Research

The process of designing, implementing, and writing a culminating project is an important part of the learning experience for a doctoral student. Although, a Research Degree (PhD) is an independent study where students must direct their own learning, manage setbacks, obstacles, and challenges as they arise. I had to undertake this key learning experience in the midst of a global crisis.

The pandemic impacted my research is multifaceted, however one major area is my research design for my third study. My PhD requires face-to-face working with people and cannot undertake my final study due to the pandemic, while social distancing measures are in place hence, I have to change track completely from a study I have plan, designed, gained approval from my supervisor and was ready to recruit participants before the pandemic hit. As the pandemic continued and I was unsure when it will be possible to have face-to-face meetings with people, I have to completely change the study's (third study) methodology.

The research was design to conduct a study on the impact of test facilitator's presence on think-aloud within usability testing. This requires face-to-face interaction with participant. I had to pause all my laboratory work and make changes to my research design in response to COVID-19 restrictions, as face-to-face interaction with participants was impossible, also, due to the lockdown, university closures and the increased risk to participants. Hence, I had to explore other factors around the use of the concurrent think-aloud and reactivity, which cause the study plans and methodology to change to interviewing usability experts. see chapter six for details.

### 1.6 Original Contribution

This research will be the first to systematically investigate the impact of task type on reactivity in CTA within usability testing and the first to systematically investigates practitioners working habit in terms of their views on reactivity and when using the concurrent think-aloud within usability testing.

Thus far, research regarding the classic think-aloud and reactivity focus on methodological differences (Van den Haak, de Jong and Schellens, 2009; Bowers and Snyder, 1990; Eger et al., 2007; Van Den Haak et al., 2003) and think-aloud instructions (Hertzum, Hansen and Andersen, 2009; Olmsted-Hawala et al., 2010; McDonald, Zhao and Edwards, 2015; McDonald, Zhao and Edwards, 2013). Few researchers have examined the impact of tasks difficulties on the CTA and reactivity (McDonald, McGarry and Willis, 2013).

While these studies provide insights on the CTA and how it could be used to evaluate digital products and provide possible recommendations regarding the CTA instruction, methodology and how participants felt about using the CTA, they did little on exploring other contributing

factors that might likely influence reactivity. Hence, there is a clear gap in the literature with regards to the exploration of other factors such as task-types that might influence reactivity. This research significance and contributions include the following:

1. It will provide practitioners with a deeper insight into the conditions that affect the validity and reliability of their test data, allowing them to make better-informed decisions about administering CTA in usability testing. This thesis will culminate in a set of recommendations that practitioners may use to guide test design.

2. Studies with the field of usability testing revealed discrepancies between Ericsson and Simon's recommendation with regards to the use of the CTA and reported practices in human-computer interaction studies (Boren & Ramey, 2000; Nørgaard & Hornbæk, 2006; Shi 2008; McDonald, Edwards, & Zhao, 2012). In addition, usability test facilitators often use instructions and probes to modify data elicitation procedures during usability experiments in order to obtain desired results. It has been argued that usability practitioners need level 3 verbalizations despite their influence on task performance.

However, usability practitioners argue that level 3 verbalisation provide them with the most useful data when identifying usability problems in a digital product such as software and website and when deriving means to resolve the problems (Olmsted-Hawala et al., 2010). Also, Study conducted by Zhao and McDonald, (2010) indicated that users do give level 3 verbalisation during usability test session even when they are not prompted to do so.

Thus, this research will provide usability practitioners with the relevant information they need to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity.

## 1.7 Approach

The studies conducted in this thesis used a single method exploration approach, a single method examines one method of think-aloud protocol and how to use the method to its best effect. Whereas a global method is when two or more think-aloud protocols (CTA, RTA and CI) are compared to find the best method to conduct usability testing.

Firstly, the rationale *for using a single method approach for this thesis is that* this thesis focuses on reactivity and its impact of the think-aloud techniques, hence, it focuses on a single method approach which is the concurrent think-aloud, rather than a global method which involves RTA and CI.

Secondly, there appear to be irregularities with regards to the finds and conclusions drawn from current academic thinking on the methodologies of the think-aloud protocol. Specifically, the CTA lacks a standard method as different studies apply different procedures, thus, a classic think-aloud might not be a classic think-aloud due to divergent practice (Gray and Salzman, 1998). As a result, a global comparison might be misleading due to lack of adopted standard of implementation which could affect the reliability of data elicited.

Thirdly, the concurrent think-aloud is widely used by usability practitioners when conducting usability testing (Boren and Ramey, 2000; McDonald, Edwards and Zhao, 2012). It is efficient when compared to RTA and it is cost-effective when compared with CI (Van Den Haak et al., 2004; McDonald et al., 2012).

Finally, one or more think-aloud protocol can be used to address the same task, however, no individual think-aloud protocol has been proven to be superior in solving all usability related problem because, every study has its unique aim and objective with different task and different artefact tested, consequently, comparing a group of study under the umbrella of CTA with studies on RTA will result in conflicting findings and bias conclusions because different study measures different criteria and cannot be generalised.

McDonald, Zhao and Edwards, (2013) emphasised that combining different think-aloud methods leads to a better understanding of usability problems, like the in case of CTA which identifies more usability issues. Whereas RTA data provides a better understanding of CTA data through reinforcement, elaboration and context of use.

## 1.8 Structure of the Thesis

This thesis encompasses the following six chapters: Chapter two, critical review of related literature, detailed discussion of the use of the classic think-aloud within usability testing and its implications regarding reactivity.

Chapter three, methodological exploration based upon past studies, detailed the experimental approach that was used to address the study's research questions, this includes the rationale for adapting to a specific experimental design. Also, each study detailed the study design and piloting.

Chapter four, first study, which is a baseline study, investigate other test factors such as task-types to gain more substantial insight into the issues that causes reactivity in the concurrent think-aloud protocol during usability testing. This study will add to a growing body of research within usability testing since it will be the first to systematically investigate the impact of tasks type on the concurrent think-aloud approaches on reactivity. It will provide practitioners with a better understanding of the conditions that affect the reliability of their test data, and it will also provide valuable recommendations to help usability practitioners guide test design.

Chapter five, second study, the study focusses on issues relating to the working habits of usability practitioners with a focus on a major aspect of divergent practice such as test facilitators' use of instruction during usability testing.

Chapter six, third study, the third study explores practitioners' views on reactivity, how they use the CTA method and the challenges they face when using the think-aloud method within usability testing in the industry.

Chapter seven, Discussion, and conclusions, reflects on the original research questions and discusses the implications of the study's findings in relation to the literature. Also, Findings obtained from the studies was discussed, the chapter concludes and presents limitations, recommendation, and future research.

Finally, chapter eight, reflection and future work, this chapter detailed a critical reflection of the entire research.

# CHAPTER TWO

## 2.0 CRITICAL REVIEW OF RELATED LITERATURE: THINK ALOUD & REACTIVITY

### 2.1 Overview

In this chapter the research background and context are presented. An in-depth exploration of the research literature about the use of think-aloud and reactivity within usability testing is reviewed. Although the think-aloud technique has been frequently employed in usability testing, it appears from a review of the pertinent literature that the significant concerns of reactivity are not thoroughly recognised. The focus of the current study is on the discussion of concurrent think-aloud, the impact of task-type to influence reactivity. It outlines a number of conceptual gaps that need to be investigated in order to better understand concurrent think-aloud and reactivity problems in usability testing.

### 2.2 Usability

The International Organisational for Standardisation (ISO) 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” (ISO 9241-11, 1998). Although, in 2010 the standard was extended with ISO 9241-210, which clearly defines experience as user’s perceptions and responses resulting from the use or anticipated use of a product which encompasses the users' perceptions, emotions, beliefs, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use as converging. According to McDonald and Petrie, (2013) the principal aim of usability evaluation is to simulate real-world product use in order to identify and rectify interface problems. Indeed, to achieve a successful design with explicit regards to effectiveness, efficiency and satisfaction usability practitioners need reliable and robust evaluation techniques to conduct usability evaluation.

There are usually four major Usability Evaluation Methods (UEMs): (1) Cognitive Walkthrough (CW) which was introduced by Peter Polson, Clayton Lewis, John Rieman and Cathleen Wharton (Lewis *et al.*, 1990). CW aims to "provide a tool for assessing the usability of a system, and assigning causes to usability problems early in the design process". This is done based on detailed specification document, mock-ups or functioning systems and it is appropriate for the development of applications where users must master a new application by learning through exploration. (2) Heuristic Evaluation (HE) which was introduced by Jakob Nielsen and Rolf Molich, (Nielsen and Molich, 1990).

It involves evaluators examining the interface of a digital product to determine its compliance with recognised usability principles. (3) The use of models; and evaluation through user participation such as the use of Think-aloud protocol. While the use of models such as the Goals, Operators, Methods and Selection (GOMS) was introduced by Stuart Card, Allen Newell and Thomas Moran (Card et al., 1983). GOMS is used to improve the efficiency of human-machine interaction by identifying and eliminating unnecessary user actions. GOMS describes the four cognitive components of skilled performance in tasks, by analysing behaviour in terms of users Goals; is used to express what the user wants to achieve; Operators are the basic actions that the user must perform to use the system; Methods are typically several ways in which a goal can be split into sub-goals and how long a task would take; and Selection rules where more than one method exists (Helander, 1988). (4) Evaluation through user participation such as the use of Think-aloud method. The subsequent section will give a brief detail of TA in usability testing.

### 2.2.1 Empirical and Inspection Methods

Empirical methods are based on capturing and analysing usage data from real end-users. Real end-users employ the software product or a prototype to complete a predefined set of tasks while the test facilitator (human or specific software) records the outcomes of their work.

Analysis of these outcomes can provide useful information to detect usability problems during the user's task completion. While Inspection methods such as Heuristic Evaluation are performed by expert evaluators or designers (i.e., they do not require the participation of real end-users) and are based on reviewing the usability aspects of Web artefacts, which are commonly user interfaces, concerning their conformance with a set of guidelines. These guidelines can range from checking the level of achievement of specific usability attributes to predictions of problems related to user interfaces.

In the Web domain, both empirical and inspection methods have several advantages and disadvantages, bearing in mind that, the majority of Web applications are developed for many different end-user profiles, empirical methods can take into account a wide range of end-users. However, the use of empirical methods may not be cost-effective since they require many resources such as a physical laboratory with computer devices for conducting experiments. Empirical methods also need full or partial implementation of the Web application, signifying that usability evaluations are mainly moved to the last stages of the Web development process.

Inspection methods, on the other hand, allow usability evaluations to be performed on Web artefacts such as mock-ups, paper prototypes, or user interface models. Also, inspection methods require fewer resources than empirical methods. However, the usability evaluation performed may be limited by the quality of the guidelines or evaluator expectations. Moreover, the interaction of real end-users is not taken into account in inspection methods.

Several studies have reported evaluations and comparisons concerning UEMs (Gray and Salzman, 1998; Hartson et al., 2001; Somervell and McCrickard, 2004). Gray and Salzman conducted an in-depth analysis of five experiments that compare usability evaluation methods intending to demonstrate that there is a need for scientific rigour in usability evaluation experiments. They claim that most experiments on the comparisons of UEMs do not clearly state or identify the aspects of UEMs that are being compared, thus producing misleading

results in an attempt to determine whether one UEM is more effective than another under certain conditions. Although the studies conducted by Gray and Salzman may be relevant in the field of HCI, there is still no well-defined research techniques that justify the studies they choose to analyse.

### 2.3 The Think-aloud Protocol

Think-aloud is one of the most used and direct techniques used to gain information about participants' thinking processes (Erisson and Simon, 1980). Hence, think-aloud is applied in different areas of research and one of the most used method in usability testing (Rubin and Chisnell, 200/ p. 204). The think-aloud method has been used in psychological study to examine the cognitive processes required in problem solving. according to Fox, Ericsson and Best, 2011) John Watson (1920) was the first to report on the use of think-aloud as he attempts to gain a deeper insight into the psychology of thinking. Duncker (1945; original German version 1935) then was among the first researcher to apply the think-aloud method in empirical studies, using it to solve mathematical problems.

Duncker made a clear distinction between think-aloud and introspection. accordingly, in think-aloud, participants "allow their activity to become verbal" instead of explaining reasons for their action. Duncker encouraged participants "not to leave unspoken even the most fleeting or foolish idea" this is to ensure that participants' verbalise all their ideas. (Duncker 1945).

Ericsson and Simon (1980) categorise verbal reports into three distinct categories. (i) time of verbal reporting; (ii) the level at which participants' think-aloud; and (iii) probing. The time of verbal reporting is a vital aspect of the think-aloud method because the aim is to obtain verbal report from the working memory (J.Karat 1997).

For the level at which participants' think-aloud, Ericsson and Simon (1980) outlined three levels of verbalisation in relation to the need of participants'' thoughts processing: (i) direct

verbalisation, this is when the information is told comparatively as it is processed in the short-term memory, and this is called level 1 verbalisation. (ii) In level 2 verbalisation, the original information is not in verbal form, hence it has to be interpreted into verbal form, a good example is an image that has to be translated into verbal form. (iii) while for level three verbalisation, participants are been instructed to procreate new principles or strategies other than just their thoughts, this makes them to filter or select information in accordance with the give instruction.

Ericsson & Simon stated that if participants' uses level one or two verbalisation, the cognitive processes remain consistent with the silent condition. They added that, level two verbalisation may show down performance. While in level three verbalisation, there is a change in participants behaviour as they focus more on information that helps them to accomplish their task efficiently. For instance, If participant are asked to describe activities that they would not otherwise pay attention to (e.g. routine actions), then it's a level three verbalisation (Ericsson & Simon 1980).

Ericsson and Simon (1980) also stated three categories for probing. In the first category participants are asked to think-aloud simultaneously with information processing. In the second category, participants' may perform their tasks silently, but the test facilitator probes concurrently based on their performance for specific information. In this instance, Ericsson and Simon suggested that general probing should be used instead of asking for specific information, if the test facilitator is still interested in information that is likely to still be in the participants short term memory. In the third category, the information is asked retrospectively, that is when participant has completed the task. If participant performed a number of tasks before the test facilitator probes, Ericsson and Simon refer to it as interpretive probing, and

are doubtful on the quality and accuracy of participants' memory of their cognitive processes during and after completing a taskset (Ericsson and Simon 1980).

### 2.3.1 Types of think-aloud methods

There are different types of think-aloud method that is been used for usability testing. We have the concurrent think-aloud; Retrospective think-aloud; Constructive interaction and the Coaching method. The concurrent think-aloud is the most popular technique among all as 85% of survey respondent select it as their most frequently used techniques compared to retrospective think-aloud and Constructive interaction (McDonald et al. 2012). In the concurrent think-aloud as the name implies user's verbalisation takes place simultaneously with their task's performance. It is primarily used to understand users' task based cognitive processes and it is both time and cost-effective McDonald and H. Petrie, (2013).

Although, the concurrent think-aloud is the most used for usability test, there are three major concerns: firstly, is artificial as users do not think-aloud when using a product in real world van den Haak, de Jong and Schellens, 2004). Secondly, depending on the think-aloud instruction, requesting participant to think-aloud may interfere with their task solving strategies (Fox, Ericsson and Best, 2011). And thirdly it is participant dependent.

The retrospective think-aloud involves verbalisation after task completion and does not interfere with participant task solving strategy. However, a major concern is whether participants are indeed able to remember everything they thought during their task performance. As it relies heavily on participant memory (Guan et al., 2006)

Apart from the single-user usability testing methods: concurrent think-aloud and retrospective think-aloud. There are less frequently employed multiple-user method such as pluralistic walkthrough which involves a group of people working together to solve a task (Bias, 1994; Van Den Haak, et al., 2006; McDonald et al., 2012).

The most employed multiple-user method is one that was originally developed in the early 1980s by Miyake (Miyake, 1982). She requested that participants should learn how a sewing machine makes stitches and made them work in two-person teams to see what they could learn. Expecting to discover to what degree the information shared by participant would improve their learning cycle. Miyake's technique is called Constructive Interaction. However, one may argue that Constructive Interaction is not a think-aloud method as it involves two users having a dialogue and the elicited data is in a conversational tone.

### 2.3.2 Think-aloud within Usability Testing

Studies, conducted by Jørgensen (1990) & Wright and Monk (1991), indicated that thinking-aloud is an effective method in user interface design as it helps product developers in detecting usability problems, mostly if the developer facilities the usability test to obtain direct feedback from participants. Think-aloud method was established in mid-90's as a principal aspect of usability testing practice (Nielsen 1993; Dumas & Redish 1993; Rubin 1994) and are considered as the "gold standard" for usability evaluation (Hornbæk, 2010). More recently, findings from study conducted by McDonald, Edwards & Zhao, (2012). Indicated that, think-aloud method is widely used by usability practitioners when conducting usability testing.

The Studies conducted by Ericsson and Simon (1980, 1984) are mostly cited as a reference for the use of think-aloud in usability testing. However, its mostly introduced without any references (Tullis & Albert 2008, p. 57). Also, in the event where Ericsson and Simon's work is referenced, the use of the think-aloud method is not applied in accordance with their instructions. For instance, studies conducted by Boren and Ramey arguably were the first to bring to light the discrepancies between Ericsson and Simon's guidelines on the use of think-aloud within usability testing and the practice of using think-aloud techniques in the industry (Boren & Ramey, 2000).

Findings from study conducted by Nørgaard and Hornbæk (2006) shows that, practically test facilitators ask more leading questions about expected problems instead of previously experience problems. Similarly, Shi reported practices of and particular challenges in using think-aloud protocols (2008). In contrast, McDonald, Edwards, and Zhao conducted an international survey study to understand how think-aloud protocols were used in a broader scale and distributed the survey to UX professional and academia (2012).

A recent study conducted by Hertzum, (2016) indicated that, participants spoke an average of 110 words per minute during a usability test session and the test facilitator who moderated the sessions spoke an average of 26 words per minute. Hence, indicating that most practitioners do not follow the established guidelines outlined by Ericsson and Simon.

Moreover, recent research has also urged the HCI community to learn more about the current UX practices in industry (MacDonald & Atwood, 2013).

### 2.3.3 Method Subjectivity & Practitioners Biases

Usability testing is a subjective process as the evaluation criteria used to assess the usability of a product are based on the evaluator's perception, experience, and knowledge. According to Kujala et al. (2011), subjective opinions may vary from one evaluator to another, leading to different outcomes and results. As a result, methods subjectivity and practitioners' biases can affect the validity and reliability of usability testing results (Hertzum & Jacobsen, 2011). One common bias is confirmation bias, where practitioners seek out information that confirms their existing beliefs and ignore information that contradicts them (Lethbridge et al., 2011). This bias can lead to a narrow focus on certain aspects of usability and a failure to consider alternative perspectives.

Another common bias is anchoring bias, where practitioners rely too heavily on initial impressions or information when making judgments (Kujala, 2011). For example, if practitioners have a negative first impression of a product, they may be less likely to identify

positive aspects of usability during testing. Conversely, if practitioners have a positive initial impression, they may overlook usability problems that are present.

Availability bias is another type of bias that can affect usability testing results (Woolrych & Cockton, 2011). This bias occurs when practitioners rely on information that is readily available to them, rather than seeking out more comprehensive information. For example, if practitioners have a limited understanding of a user group, they may rely on stereotypes or assumptions rather than conducting thorough research.

Overconfidence bias is a bias where practitioners overestimate their own abilities or the accuracy of their judgments (Hertzum & Jacobsen, 2011). This bias can lead to a failure to identify usability problems or a failure to recognise the limitations of their own expertise. Additionally, practitioners may be less likely to seek out feedback from other professionals or users if they are overconfident in their own abilities.

Finally, selection bias is a type of bias where practitioners select participants or data that supports their existing beliefs or hypotheses (Kujala, 2011). This bias can lead to a skewed sample that does not accurately represent the user population. For example, if practitioners only recruit participants who are already familiar with a tested product, they may overlook usability problems that are present for new users.

#### 2.3.3.1 Addressing Methods Subjectivity and Practitioners Biases in Usability Testing

To address methods subjectivity and practitioners' biases in usability testing, it is important to implement rigorous testing processes that are designed to minimise these biases. One approach is to use objective measures of usability such as task completion time or error rates

can help to reduce the impact of subjective judgments (Lethbridge et al., 2011). This type of metrics was used within the studies in this thesis.

Another approach is to conduct thorough research on user groups and user needs prior to testing (Woolrych & Cockton, 2011). This can help to reduce biases such as availability bias and selection bias, by ensuring that practitioners have a comprehensive understanding of the test participants. Also, incorporating user feedback throughout the design process can help to reduce biases such as overconfidence bias and anchoring bias, by ensuring that practitioners are open to alternative perspectives and willing to make changes based on participant's feedback. More detailed discussion on the usability subjectivity and evaluator's effect is discussed in section 2.8: the evaluator effect.

## 2.4 Empirical Studies of Think-aloud within Usability Test

Think-aloud (TA) protocol is one of the fundamental tools and has been widely used by usability practitioners when conducting usability testing (Boren and Ramey, 2000; McDonald, Edwards and Zhao, 2012). The primary aim of TA is to gain insight into users' task solving strategies and behaviours by obtaining a report on the user's experience when interacting with a design in order to identify usability problems and possible suggestions for improvement (Van Den Haak, De Jong and Jan Schellens, 2003; Olmsted-Hawala *et al.*, 2010; McDonald, Zhao and Edwards, 2013a). TA protocols encompass: The Concurrent Think-aloud (CTA), Retrospective Think-aloud (RTA) and the Constructive Interaction (CI).
The next section will discuss each of the think-aloud protocols in detail together with their individual strengths and weaknesses.

### 2.4.1 Concurrent Think-aloud (CTA)

In CTA the users' verbalisation takes place simultaneously with their task performance; the CTA is primarily used to gain access to users' direct report of their tasks performance. This

should result in a more complete overview of usability problems encountered in addition to the observable problems that may be extracted from interface behaviours. Verbalisations may reveal any doubts, irritation, surprise or other feelings that arise during the process (Van Den Haak, De Jong and Jan Schellens, 2003; Cooke, 2010). CTA also show how different uses make use of a product which might not directly lead to a usability problem, however may mirror some aspect of the use of a product. This is what Hertzum, (2010) referred to as the images of usability. Thus, might be helpful to usability practitioners when recommending improvement for a design.

According to McDonald, McGarry and Willis, (2013), CTA is used to understand users task based cognitive processes both in usability testing and in HCI. It is also CTA is also advantageous with regards to time and cost of conducting usability experiment (Van den Haak, de Jong and Schellens, 2004).

Ericsson and Simon, (1980, 1993) established guidelines to ensure the validity of CTA data elicitation, these guidelines include: (i) natural instruction that does not request specific types of information (ii) a practice session and (iii) a neutral “keep talking” prompt with no addition additional evaluator probes that need to be adhere in order to ensure validity. When all three guidelines are used collectively there are refer to as the “Classic TA protocol”. They also categorised participants’ verbalisations into three different levels: Level 1 verbalisation as the most reliable because they are direct data elicitation of user behaviour from the short-term memory; Level 2 verbalisations are subject to an intermediary process in which users must transform abstracts concepts into words; while level 3 verbalisations are considered invalid as it requires additional cognitive processing of information from long term memory. See table 2 for detail.

**Definition of Verbalisation Levels**

| Simon & Ericsson Verbalisation Definition | Simon & Ericsson Verbalisation Example | Example from Current Study |
|---|---|---|
| Level 1<br>Verbalisation is simply the verbalisation of conscious thoughts where participants do not need to make any effort to communicate what they are doing and how they do it. | Reading a sentence aloud | So we want a twin room with bed and breakfast and upon arrival we require airport transfer on a Luxury Eco car |
| Level 2<br>Verbalisation involves the conversion of content in short term memory into words. | Conveying an image or object in words and does not require additional intellectual work | So okay store locator and I will go for North Shields and I will type North Shields,<br>So North Shields Thomas cook and I want the postcode it said NE29 6QF |
| Level 3<br>Verbalisation requires participants to explain their thoughts, ideas, notion or motives. | Explaining and reflecting on personal experiences, feelings and events. | I will say that Thomas cook is been dodgy because their claims positions are running away, although this is very simple I think I did well although the site the links are broken. |

*Table 2: Simon & Ericsson's verbalisations categorises*

Even though Ericsson and Simon's model was developed within the field of cognitive psychology, it has since provided the rationale for collecting verbal data in other fields such as: Engineering and usability testing (Denning *et al.*, 1990; Pressley and Afflerbach, 1995). CTA protocols are a widely used method by usability practitioners to conduct usability test, however CTA methodological implementation in usability literature and the work habits of practitioners do not conform to the theoretical foundation established by Ericsson and Simon: Protocol Analysis: Verbal Reports as data, due to the fact that usability evaluators often focus on collecting level 3 data. A major aspect of the divergent practice is evaluators' intervention during CTA usability tests by using probing questions and instructing participants' to comment on specific instances in order to obtain desired results (Boren and Ramey, 2000; Nørgaard and Hornbaek, 2006; McDonald, Edwards and Zhao, 2012). According to Ericsson and Simon Framework, any verbalisations prompted by evaluator is categorised as level 3 verbalisation and considered as invalid due to their subjective content and access to long-term memory (Ericsson and Simon, 1980, 1993).

Also, a meta-analysis conducted by Fox, Ericsson and Best, (2011) indicates that, a deviation from these guidelines often induce reactivity: change in cognitive processes that goes beyond the scope of the working memory or tasks explanation (Van Den Haak, De Jong and Jan Schellens, 2003; Fox, Ericsson and Best, 2011). However, the evaluators who elicit them may not consider them to be level 3 data and might as well not consider them as invalid data.

Boren and Ramey, (2000) reviewed the ways in which actual usability practice diverges from Ericsson and Simon's established guidelines and highlighted the difference between what practitioners do and what Ericsson and Simon's theory would allow. They reported that practitioners often give participants' instruction that are contrary to the guidelines established by Ericsson and Simon, instead of using the simple 'keep talking' reminder they make use of probing intervention. Boren and Ramey, (2000) also, argued that other framework such as Speech Communication Theory, might be more suitable in handling such issues that arise within usability testing.

CTA has been known to cause reactivity, according to Van den Haak, de Jong and Schellens, (2004) reactivity sometimes occurs among participants as a result of having to combine thinking-aloud with task performance. Also, this change in cognitive processing: reactivity, is problematic due to its impact on task performance either positively or negatively. The former result to an improved performance which can be referred to as the self-explanation effect which might lead to failure in detecting interaction problem (Chi *et al.*, 1994; Nathan, 1994) and the latter, a decline in performance which can be referred to as verbal overshadowing which might lead practitioners to identify and potentially rectifying usability issues that are unlikely to be encountered by end users' in real world (Chin and Schooler, 2008).

### 2.5 Reactivity of the think-aloud method

The issue of reactivity during CTA data elicitation within usability testing is a subject of much debate, emphasis is placed on the fact that when participants are asked to perform certain

tasks and think aloud simultaneously, they might experience difficulty in verbalisation and also perform the task distinctively due to their combined task at hand being too high and could result in reactivity (Russo, Johnson and Stephens, 1989). This is problematic because it may affect the way in which task performance is been measured and may also lead to poor usability problem detection (McDonald, Zhao and Edwards, 2013).

Ericsson and Simon, (1980, 1993) stated that, "the accuracy of verbal reports depends on the procedures used to elicit them" and reactivity will occur when the established procedure is neglected. Likewise, Hertzum, Hansen and Andersen, (2009) highlighted that the act of thinking aloud alone is unlikely to cause reactivity, except for methodological irregularities.
To gain a deeper insight into the issue of reactivity within the context of usability testing, I will discuss studies where no modification to the procedures established by Ericsson and Simon has been made, that is studies that compares CTA with silence, then I will look at when the guidelines is been modified and then when probes are implemented

### 2.5.1 Reactivity: Evaluation of studies with no procedural modification

Most usability testing with evidence of reactivity tends to be conducted using global think-aloud, that is experimental comparison between CTA and RTA where the formal encompasses participants' verbalising concurrently while performing a task and the latter involves participants working in silent and verbalisation take place after task completion.
Van Den Haak et al., (2003) conducted an experiment that compares CTA and RTA for a usability test of an online library catalogue. The type of tasks involves fact finding where participants was asked to look for specific information on the catalogue. Results indicated that CTA method caused reactivity as the tasks of concurrently verbalising thoughts causes the participants to make more errors in the process of task performance and to be less successful in completing tasks compared to those working with RTA.

The most plausible explanation for this observation regarding the CTA method lies in the participants' workload that is, the level of difficulty of the tasks given to the participants may have been a crucial factor in this study, although there is always a possibility that a problem detected in a CTA usability test is partly caused by the adopted method which seems not to be in accordance with Ericsson and Simon, (1980, 1993) established guidelines.

Also, study conducted by Eger et al., 2007 aimed to examine the validity of retrospective verbal reporting cued by eye movement replay in a web-based usability context and also to assess the reactivity effects associated with thinking aloud. Their findings indicated that, the eye-cued method identified more usability problems than the think-aloud or screen-cued methods and fewer participants completed the search task on the Think-aloud condition, indicating the reactivity of the technique. Thus, the results demonstrate that retrospective methods cued by eye-movement data can be more insightful, beneficial, less-reactive and more informative to usability evaluator than a conventional think-aloud protocol.

A study conducted by Bowers and Snyder, (1990) compared concurrent and retrospective verbal elicitation techniques to show whether a large monitor would be advantageous to the user compared to a small monitor for windowing tasks and also if task type would affect these advantages. The study makes used of a mixed factorial design, using a between-subject treatment of verbal protocol with 48 participants with a well detail experimental procedures for repetition.

Findings indicated that, due to the level of task difficulty, participants in the concurrent condition were forced to give verbalisations that requires further cognitive processing, thus inducing reactivity. Also, CTA participants did give very low-level processing: in this context, difficulty to verbalise their thoughts in accordance with the task performance which also result in performance degradation. Although, there were more verbalisation for high difficult tasks than for medium and low difficult task which reflects the number of steps it took to complete a

task, however there were no significant differences in task completion time between the two think-aloud conditions.

Van den Haak, de Jong and Schellens, (2009) conducted a study to gain a more substantiated picture whether the usability method that was previously employed for the evaluation of an earlier municipal website (Van Den Haak, et al 2007) would reveal the same or different results when applied to a municipal website with a different information architecture.
Findings indicated that, participants using the CI method find it more difficult to perform reading asks than navigation tasks due to the fact that tasks type associated with reading involve an inherently individual process while tasks type associated with navigation involve physical actions that are visible to both participants and can thus be discussed more easily.

Similarly, RTA participants verbalise more problems when their task type involves navigation than when their tasks type is associated with reading. While CTA participants might experience more or less reactivity depending on the type of task. Overall, CTA, RTA and CI were comparable in terms of result they produce and there was no difference regarding task completion times and number of tasks completed successfully.
However, findings indicate that the three evaluation methods might work differently depending on the nature of task performed and the information architecture of the website that is been tested.

Evidence from the above studies seems questionable and inconclusive with some studies showing reactivity and other showing non reactivity of CTA. This suggests other contributing factors might be influencing performance in addition to the think-aloud protocol or could it be task type? However, some of the studies do not report CTA data elicitation procedure in sufficient details to verify if there are in accordance with the stringent guidelines established by Ericson and Simon.

This section has review studies with no procedural modification to CTA, however, one cannot draw a valid conclusion due to irregularities of the results obtained from different studies to whether CRT on its own causes reactivity or if it was the elicitation process, hence the next section will review studies with procedural modification to CTA to gain a deeper understanding into reactivity occurrences during usability testing.

### 2.5.2 Reactivity: Evaluation of studies with procedural modification

In usability testing, evaluators often use instructions and probes to modify data elicitation procedures during usability experiments in order to obtain desire results. This modification is what Ericsson and Simon referred to as level 3 verbalisation and considered it to be invalid. In this section I will review studies with instructions modification and when probes are implemented.

#### 2.5.2.1 Instructions within Usability Testing

The use of think-aloud (TA) in usability testing has continues to be an important and active area of research with a much focus on issues relating to the working habits of practitioners which do not conform to the theoretical foundation established by Ericsson and Simon for collecting verbal data during the use of CTA within usability test.

These guidelines include: (i) natural instruction that does not request specific types of information (ii) a practice session and (iii) a neutral “keep talking” prompt with no addition additional test facilitators probes that need to adhere in order to ensure validity (Ericsson and Simon, 1993, 1980). According to Ericsson and Simon Framework, any verbalisations prompted by the test facilitator is categorised as level 3 verbalisation and considered as invalid due to their subjective content and access to long-term memory.

The concurrent think-aloud utility involves different types of instruction when used in usability testing, these are: Classic think-aloud, relaxed think-aloud and explicit think-aloud. When all three guidelines that was established by Ericsson and Simon as mentioned above are used

collectively there are referred to as the “Classic TA protocol”. The relaxed think aloud also known as interactive think-aloud constitutes a relaxation of the think-aloud protocol, in which users are requested to verbalise their thoughts by providing a running commentary on their actions and are prompted for their current thoughts as well as their reflections and actions (Hertzum et al., 2015). While explicit think-aloud also known as explicit instructions involves giving users’ direct instructions to explain their navigation decisions.

Studies have documented divergent practice in the use of think-aloud instructions and test facilitators interventions (Boren and Ramey, 2000). This is problematic because they might sometime threaten test reliability by inducing reactivity: a change in task performance which makes a usability test no longer a representation of real-world use.

Studies have compared both classic and relaxed think aloud or explicit instructions, however, none has compared CTA; Explicit instruction and Silent working with Fact, and assessment tasks to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity. For instance, Hertzum et al., (2015), investigated verbalisation in usability test by comparing participants verbalisation in moderated and un-moderated test during relaxed think-aloud, findings indicated that the verbalisations made by moderated and un-moderated participants were similar in content and the main difference being a higher percentage of high relevance verbalisations by un-moderated participants with action description, system observation, and user experience were the most frequent categories.

A study conducted by McDonald and Petrie, (2013) investigated whether the classic think-aloud and a think-aloud with an explicit instruction led to different task solving performance compared to silent working. The result shows that for classic method there was no impact on task performance, however, the explicit instruction led to an increase in within and between page navigation and scrolling activity, as a result, may lead to changes to what users do at the interface.  Similarly, Zhao et al., (2014) compared the classic think-aloud and an explicit

instruction. The classic think-aloud instruction was in accordance with the guidelines given by Ericsson and Simon, while the explicit think aloud requested participants to verbalise their expectations, surprises, delights, confusions, frustrations and content that is relevant to the user experience. All other test conditions were kept similar and the only interaction between the test facilitator and participants was a reminder to "keep talking" if participants fell silent for 15-20 seconds. The study adopted a between subject's design to avoid possible transfer effects and the dependent variables were task performance data, participant mental workload measured with NASA Task Load Index, utterance data and usability problem data such as number, severity, types and source.

Findings indicated that there was no difference in task performance and explicit instruction did not lead to reactivity. In terms of participant utterances, there was no difference in the number of utterances made from both conditions with the predominant utterances related to procedural descriptions and reading activities. However, participants on explicit instructions condition reported an increased in cognitive workload, also they assessed their own behaviour as being more focused on finding problems than the participants in the classic condition. Hence, their conclusion suggested that, although the explicit instruction did not lead to reactivity, however, its impact may influence participants to be hypersensitive to interface issues to comply with the instruction.

McDonald et al., (2013a) investigated whether an explicit explanation-based think-aloud instruction leads to differences in navigation performance over the classic think-aloud method with silent performance. Findings indicated that, in terms of the impact of tasks difficulty, CTA participants completed fewer tasks successfully and that the classic think-aloud did not change participant's behaviour, although it increased participants' ratings for effort and frustration. Whereas for easy tasks there were no differences in task performance among participants' when compared to the explicit condition. In terms of verbal utterances both conditions were dominated by procedural descriptions, with the explicit think-aloud having more explanatory utterances.

Cooke, (2010) study addresses the use of think-aloud protocols in usability test settings with respect to users' verbalisation accuracy, verbalised content, and what do users' eye movements reveal about their behaviour when they are silent. In terms of verbalisations findings indicated that verbalisation for CTA was mainly procedural in nature with participants often reading from the screen. Also, Zhao and McDonald, (2010) compared the classic and a relaxed think-aloud with the aim to explore the impact of think-aloud style on the nature of the utterances produced by participants and the usefulness of those utterances for usability analysis. Findings indicated that the interactive think-aloud led to the production of more utterances than the classic think-aloud, such utterances categories include problem formulation, causal explanation, user experience and recommendation which could be used in usability problem analysis. However, no significant difference was found between interactive and classic think-aloud for utterance categories such as action description, reading and task confusion.

Wright and Converse (1992) investigated the impact of concurrent verbalisation on task performance during usability testing. They compared two groups of participants to solve file management tasks, one group worked in silent and the other provided a concurrent think-aloud while providing explanations for their actions; that is, level 3 verbalisation. If participants in the CTA condition were silence for more than 30 seconds or issued a command without giving an explanation they were prompted for their thoughts and reasons, thus using both instructions and probes.

Results indicated that, participants in the CTA condition committed fewer errors, consumed less task time and performed better than participants' in the silent condition. However, it is debatable as one might raise a question if the findings should focus merely on the elicitation procedure given the fact that the CTA condition also included evaluator interventions and probes. These results were extremely important in revealing a potential method bias in

usability tests; thereby confirming Ericsson and Simon's contention that providing explanations will result in level 3 verbalisation and possibly performance improvement.

McDonald, Zhao and Edwards, (2013) investigated the benefit of collecting both CTA and RTA data in the same usability test, using 10 participants, four individual task was performed on a university intranet website. The CTA condition was conducted in accordance with Ericsson and & Simon's recommendations on CTA data elicitation. While on the RTA condition the evaluator grant users' access to the site instead of watching the video playback of their task. They also eliminate the use of prompt rather they make use of only acknowledgment token, i.e., "uh-huh, mm-hmm". Finding indicated that, the CTA yield useful data indicating when participants stray from accurate task solution and was non-reactive. While the RTA participants yielded more insight into issues recovered by the CTA participants. The significant of this finding is that a better understanding of usability problem data can be extracted by using a dual verbal elicitation procedure in the same usability test.

McDonald and Petrie, (2013) investigated the impact of think-aloud instructions on task performance by comparing CTA and a think-aloud with explicit instruction with silent. The CTA makes use of neutral instruction which is in accordance with Ericsson and Simon guidelines while the ETA makes use of explicit instruction. The study makes use of a within subject design with 8 participants, findings indicated that, the classic method had no impact on task performance; however, it contributed to an increase in effort and frustration. While the explicit think-aloud had no impact on task success and time on task but did yield some differences in task performance, suggesting that the cognitive process to verbalise in accordance with specific instructions increase task difficulty as it led to an increase in within and between page navigation.

Also, the result indicated no evidence of reactivity. However, the behavioural differences in the explicit condition suggest an increase in mental workload which might induce reactivity depending on the type of task.

#### 2.5.2.2 Probes within Usability Testing

Hertzum, Hansen and Andersen, 2009 conducted a study to compare classic think-aloud and relaxed think-aloud with silent condition. The aim of the investigation was to find out if thinking-aloud causes participants in usability evaluations to behave differently and experience a different level of mental workload compared to performing in silence. The relaxed think-aloud encompasses level 3 verbalisation and according to Boren and Ramey, (2000) corresponds to how think-aloud is commonly employed in the context of usability evaluation.

Findings from this study revealed that, the classic think-aloud have little effect on participant behaviour and cognitive workload, apart from elongating the task which was obvious in assessment task type than in fact finding task type for both the classical and the relaxed think-aloud conditions. Contrarily, the relaxed think-aloud affected participant behaviour with regards to longer task completion time, more navigation traversal and a higher mental workload. Hertzum, Hansen and Andersen, 2009 also added that, the relaxed think-aloud approach tends to threaten test validity indicating that this approach commonly employed by usability practitioners may be reactive and is not a valid method of data elicitation within the context of usability testing.

Olmsted-Hawala et al., (2010) carried out a comparison on three different types of TA protocol: a traditional TA, Speech-communication-based protocol: think-aloud following the speech communication theories and Coaching protocol; think-aloud with active intervention, with a silent condition as control which is like the relaxed think-aloud applied by Hertzum, Hansen and Andersen, (2009). The aim of the study is to provide practitioners with a better understanding of the strengths and weakness of the different variants of think-aloud protocol. The study makes use of a between-subject design with 80 participants that were randomly assisted to one of four conditions. The experimental procedure for the traditional TA was in

accordance with the guidelines given by Ericsson and Simon i.e., no probing words beyond "keep talking.

Findings indicate that, participants in the coaching condition were more successful than participants in the other two conditions: The use of active intervention promote task performance over Classic and Speech-communication conditions for both accuracy and satisfaction respectively. Finally, there was no difference among the conditions in terms of efficiency and found no reactivity.

Also, McDonald, Zhao and Edwards, (2015) conducted a study to compare two concurrent think-aloud approaches: the classic think-aloud and an Interacting think-aloud (ITA) with respect to task performance and usability problem extraction. Findings indicated no differences in the number of successfully completed task, Although ITA led to the detection of more usability problem and a greater number of causal explanations. However, the elicited data from ITA indicated low severity problem and also prolong the test session. Evidence from literature indicates irregular findings, as some studies indicate reactivity while others do not indicate reactivity.

#### 2.5.2.3 Demonstration of A Think-Aloud Practice Session

While think-aloud approaches have their origins in cognitive science, it is necessary to consider the relationship between thought and words. it is helpful to return to Ericsson and Simon (1980) seminal study, Verbal Reprots as Data where the theoretical foundation of think-aloud approaches and similar "introspective" analysis strategies was emphasised. Their philosophy is founded on a distinction between working memory, in which concurrent thought occurs verbally, and long-term memory, in which some of the thoughts from working memory are ultimately processed, but not always in words.

The aim of think-aloud study is to provide the author with insight into the working memory mechanisms, but there are other challenges that researchers must be mindful of. such challenges involve the awareness that only information that is "heeded" or "heard" is stored in working memory. Since working memory has a finite power, this knowledge is only retained for a short time before being superseded by new thinking patterns. Hence, only verbal reports given shortly after a thought process can be said to reliably represent cognitive thought, and researchers must rely on the participants' "immediate consciousness," rather than delayed reasons for their behaviour (Charters, 2003; Pike et al., 2014).

Indeed, Ericsson and Simon advocated for the use of think-aloud practise session prior to data collection for two reasons: firstly, think-aloud can be unnatural and may be uncomfortable for participants, so a practise session can help participants become acquainted with the procedure. secondly, practice appears to teach participants to verbalise in accordance with the general think-aloud guidance, resulting in the development of level 1 and level 2 verbalisations rather than level 3 verbalisations. Furthermore, Ericsson and Simon, (1993) recommended that participant should be retaught and given extra warm-up practice which will enable their verbalisation to be compliant with the general guidance, they also stress on a simple practise. For instance, a study conducted by McDonald, Zhao, and Edwards,(2013) to examined the usefulness of combining a concurrent and retrospective think-aloud within the same usability test uses a practice session. Participants were instructed to think aloud when looking up the definition of libretto in a dictionary and then disassembling and reassembling a ball point pen. However, in a previous study McDonald et al. (2012) found out that a practice session is not often used by practitioners during a usability test. although, research within the field of usability has not focus on its impact either positively or negatively on the reactivity of think-aloud.

## 2.6 Tasks and tasks-types within Usability Test

Usability testing is a task-based approach and tasks are meant to model real world use of a design and how people interact with a design on their own in order to understand the functional and non-functional aspect of a design. (Elling, Lentz and de Jong, 2012). Thus, by requesting users to perform actionable tasks, the think-aloud utility enables evaluators to gain qualitative insights into what is causing users to have trouble when participants attempt different tasks-type such as: (i) fact finding which involves search tasks to find out from users what a design requires, assessment tasks to assist usability evaluator in discovering usability issues during usability testing. (ii) assessment tasks: to assess the usability of an artefact from tasks that requires users to make judgement and personal opinion (Van Den Haak, De Jong and Jan Schellens, 2003; Van den Haak, De Jong and Schellens, 2007); and (iii) sense-making tasks: to assess users specific information needs with regards to the goal of a design by understanding and applying the relevant information to the described scenario in order to answer task related specific questions, search for better information and filter out undesired information (DiMicco and Millen, 2008, Elling, Lentz and de Jong, 2012).

Also, users are faced with the basic task of making sense of what they see when encounter with an unfamiliar design for the first time, thus how information is organised, presented, integrated and controlled directly affects how easily users will understand and analyse such design (Russell et al. 1993).

Research conducted by Van Den Haak et al. (2009) on Municipal website evaluation claims that, the information architecture of a website and different task types could potentially affect the working of think-aloud methods. They highlighted that, participants using the Construction Interaction (CI) method may find it more difficult to perform reading tasks than navigation tasks due to the fact that tasks type associated with reading involve an inherently individual process while tasks type associated with navigation involve physical actions that are visible to both participants and can thus be discussed more easily.

Similarly, RTA participants verbalise more problems when their task type involves navigation than when their tasks type is associated with substantial reading. Also, CTA participants might experience more or less reactivity depending on whether their task type involves navigation or reading. Findings indicate all methods (CTA, RTA and CI) were comparable in terms of result they produce and there was no difference regarding task completion times and number of tasks completed successfully and with the indication that the three evaluation methods might work differently depending on the task type and nature of the website (Website information architecture) that is being tested.

To gain a better understanding to the impact of task on CTA, study conducted by (McDonald, McGarry and Willis, 2013b) on the relationship between think-aloud instructions, task difficulty and performance with the aim to investigate If an explicit explanation-based think-aloud instruction leads to differences in navigation performance over the classic think-aloud method. Findings indicated that, for the low difficult tasks, there was no difference in task success however, for complex tasks there were differences, as participants in the classic condition completed fewer tasks successfully and engaged in more link traversals. This indicates that, when information scent: the strength of a cue, is low task difficulty will be increased and the task will be more cognitively demanding as a result might induce reactivity.

Similarly, Taylor and Dionne (2000) stated that, certain type of tasks can influence verbal reports due to the fact that simple tasks are processed by well learned routine which the short-term memory do not pay attention to or take note of due to its simplicity, it is often done automatically. Thus, verbalisation of such tasks may not provide quality data.

Furthermore, Van Den Haak (2003) opines that, tasks should be designed in such a way that they could be carried out independently with equal difficulty in order to avoid participants from been stuck after one or two tasks. Hence, this research will provide greater understanding of

the relationship between task-types and TA approaches and will be relevant to usability practitioners in their attempt to make effective decision in choosing the appropriate method when conducting usability testing.

### 2.6.1 Task Derivation

One of the most effective techniques to identify usability problem is to observe users while they perform series of tasks. When the representative participants attempt realistic tasks, one tends to gain qualitative insight into what causes users to have problems when interacting with an interface. Its implication is that it helps practitioners to recommend possible improvement of the user-friendliness of the tested product.

According to (Redish and Dumas 1999, p.160) usability testing is a sampling process, thus, one cannot test every possible task users can perform with a product. Hence, they suggested four major areas when selecting tasks, these includes tasks that probe potential usability problems; tasks suggested based on concern and experience; tasks derived from other criteria and tasks that users will do with the product. Similarly, (Dumas and Fox, 2009, p.233) highlighted three major areas where tasks should be structured around during the formulation of tasks for usability testing. These includes: (i) important tasks such as frequently performed tasks and basic tasks that involve the core functionalities of a product (ii) tasks that involve areas where the test evaluator preserved users might have difficulties and (iii) tasks that explore the product navigation and information architecture. They also added that tasks should be structured around business goals, product re-design and new features.

In relation to this study, all four criteria were put into consideration during the task's formulation. For example, tasks were formulated in such a way that they make sense to the average user by avoiding technical terms and adopting the use of simple languages and the

set of tasks were structured around the major activities' users will normally carry out on their daily basis when using the product.

### 2.6.2 Tasks Scenarios

Redish and Dumas, (1999) suggest the use of scenario as it makes a task more realistic and because it eliminates the artificiality of the tasks by telling a reasonable and very short story. Similarly, (Rubin and Chisnell, 2008, p.125) "Describes task scenarios as adding context, participants' rationale and motivation to perform task". They also added that the closer a task scenario to reality the more reliable the test results.

According to Nielsen, (1994) in order to engage users, a good task scenario should be realistic and actionable and should also avoid tasks clues. Also, he suggested that test evaluator should allow participants to ask a question relating to the tasks description in order to minimise the risks of tasks misinterpretation.

Hence, in relation to this study tasks were formulated in such a way that they tell a reasonable and short story and do not give clues to participants by avoiding the use of unique words that is used in the tested product, due to the fact that participants usually scan the product in search for these related words. The implication of using related words will not only bias the test results, it will also limit the number of usability problems that will be detected.

### 2.6.3 Characteristics of the Taskset

Nielsen, (1994) Suggested that the first task in a usability test should be very simple in order to increase the user confidence, boost morale and to guarantee the user an early success experience. The task set are ordered in such a way that they are independent of each other, reason been to enable flexibility in cases where changing the order of the tasks is required and also to allow continuity and progress to the next tasks in cases where participants decide

to abandon a task and to ensure that all tasks do not lead participants to the solution of other attempt tasks.

Boren and Ramey, (2000) suggest that each task should end with a statement that tells participants when a task is complete in order to avoid misunderstanding in situations where participants mistakenly think a task is completed when it is not actually complete. They added that a practical implication for not indicating when a task is complete will be usability test facilitator intervention that could alter participant's normal task flow. Hence, with regards to this study, all tasks were formulated with an indication for participants to know when a task is complete.

Overall, during the formulation of the task two aspects were practically considered. Firstly, site inspection and product walkthrough were carried out to work out the site navigations. Attention was given to the area where participants might experience difficulties. Secondly, a considerable attention was given to the main features and functionality of the site, thus a clear goal for the tested product was established. Thirdly, I ensure an appropriate level of details in order not to give participants clues and step to accomplish tasks. Fourthly, all tasks were known-items tasks: this implies that all tasks require participants to search for information that the test facilitator known to exist in the product (Kim, 2001).

Finally, all tasks should have a correct solution which is used as a way of accounting for the correctness of a task by developing a means to measure whether or not users accomplished the tasks. Thus, all tasks solution has to be written by the participants on a provided answer booklet, this implies that the correctness of each task can be easy measured both by the participants' and the test facilitator.

## 2.7 The Evaluator Effect

### 2.7.1 The impact of test facilitator's presence in usability testing

In a recent study conducted by Hertzum, Molich and Jacobsen, (2014) on the evaluator effect. According to the findings of this study, evaluators can find different sets of usability problems even while analysing the same usability test sessions. To gain a deeper understanding it is helpful to go back to one of the earliest social psychological studies on facilitation: tendency to perform tasks better or faster in the presence of others and social inhibition: the tendency to perform tasks more poorly or slower in the presence of others. Hazel Markus (1978) gave study participants a simple task (putting on and tying their shoes) as well as a more complex task (putting on and tying a lab coat that tied in the back). Participants in the study were asked to complete both tasks in one of three social settings: (a) alone, (b) with a colleague present who was watching them, or (c) with a colleague present who was fixing a piece of equipment in the corner of the room without looking.

Overall, Markus found that the difficult task was being done more slowly than normal. However, she discovered an interaction effect, in which when a colleague was present in the room, the participants performed the easy task faster but the complex task slower. Furthermore, it made no difference whether the other person was paying attention to the performance or was simply in the room doing something else, the mere presence of another person nearby affected performance.

Similarly, in the context of usability, Held and Biers (1992) were among the first to investigate the influence of test facilitators presence in usability testing. Their study sought to assess the impact of evaluator intervention, task structure, and user experience on users' subjective evaluations of software usability. The study used a two-by-two factorial between-subjects design with two levels of Evaluator Intervention (Intervention vs. Non-Intervention), two levels

of Task Structure (Guided-Exploration vs. Standard Laboratory), and two levels of User Experience (Guided-Exploration vs. Standard Laboratory) (Novice, Experienced).

The main finding was that both user Experience and evaluator Intervention influenced the user's subjective impression of the software. Experienced users rated difficult-to-use word processing features as more difficult to use under the intervention condition than under the non-intervention condition. This difference was not significant for novice users.

Yeo (1998) conducted a study to identify cultural factors that may affect results of usability evaluation techniques. Initial findings indicated that power distance was a significant possible cultural factor: A test user with a higher rank than the experimenter made more negative comments about the product than a test user with a lower rank. also, users would be more hesitant to provide negative feedback if they believed the evaluator was of higher rank because they lacked a task-focus orientation and hoped to develop a positive relationship with the higher-ranking evaluator.

In another study, Yeo (2000) conducted an exploratory study to examine Malaysian participants verbalisation recorded in think aloud and interview sessions. It was anticipated that Chinese users would consider the feelings of the moderator and refrain from making negative comments about the evaluated system. However, the findings indicate that, rather than being excessively polite to the moderator, most users concentrate on the test tasks and take on the role of assisting in the discovery of possible usability issues.

For example, a study conducted by Riihiaho, (2014) examined the effects of relaxed thinking aloud and the presence of a test moderator, findings from the study indicated a significant effect of the test facilitator’s presence is found in the users' subjective rating, as participants who carry out task performance in the presence of a test facilitator rate the system preferences significantly higher than participants performing alone (Riihiaho, 2014).

Also, Eger et al., (2007) examined the validity of retrospective verbal reporting cued by eye movement replay in a web-based usability context and also assess the reactivity effects associated with thinking aloud. They found that the impact of the experimenter's presence had a negative effect on the participants during their think-aloud reporting, with responses indicating that the moderator's presence feels more unpleasant when thinking aloud is required.

Sonderegger and Sauer (2009), examined how situational factors such as observers in usability tests affect the test results. The study uses a 3 x 2 mixed experimental design. Three conditions compared are: no person present, presence of facilitator and presence of facilitator and two non-interactive observers. Although task difficulty was controlled as a between-subjects variable (low vs. high). Data on performance, subjective measures, and physiological parameters (such as heart rate variability) were collected.

The findings revealed that the presence of non-interactive observers during the usability test caused an increase on stress level, elongated the time spent on tasks and decreased overall performance. The presence of a facilitator (i.e., a participating observer) also influenced the test participant's emotional state as participant who perform tasks with no person present rate their emotions more positive than the others. Although, users who performed alone rated their emotions more favourably, the researchers discovered that a moderator who can establish a good rapport with the test participants could also increase their performance (Sonderegger and Sauer, 2009).

Hertzum et al., (2015) conducted a study which compares moderated and unmoderated test sessions, they found that the unmoderated participants made a higher percentage of high-relevance verbalisations, an average of 21% compared to 11% for the moderated participants. Findings from McDonald and Petrie (2013) interview data shows that four out of eight participants indicated that their increased persistence was more to do with trying

to impress the test facilitator. Also, Grubaugh et al. (2005) also found higher error rates in usability testing when the laboratory set-up was more intrusive in terms of monitoring equipment used.

### 2.8 Plausible Reasons for The Divergent Use of The CTA within Usability Testing

According to Ericsson and Simon Framework, any verbalisations prompted by the test facilitator is categorised as level 3 verbalisation and considered as invalid due to their subjective content and access to long-term memory. Also, Ericsson and Simon, (1993, 1980) categorised participants' verbalisations into three different levels: Level 1 verbalisation: as conscious thoughts where participants do not need to make any effort to communicate what they are doing and how they do it i.e. reading a sentence aloud, they suggest is the most reliable because they are direct data elicitation of user behaviour from the short-term memory; Level 2 verbalisations: as the conversion of content in short term memory into words i.e. conveying an image or object in words and does not require additional intellectual work and level 3 verbalisations are considered invalid as it requires additional cognitive processing of information from long term memory.

Another major aspect of divergent practice is test facilitators' intervention during CTA usability test sessions by using probing questions and instructing participants to comment on specific instances in order to obtain desired results and factors relating to the impact such intervention might have on the elicited data in terms of validity (Boren and Ramey, 2000; Hertzum et al., 2009; Nørgaard and Hornbaek, 2006). Further, a significant factor to consider is the fact that users' do give level 3 verbalisation during usability test session even when they are not prompted to do so (Zhao and McDonald, 2010). This could be as a result of their perceived awareness that the essence of usability testing is to improve the tested product and facilitate insight problem-solving. Also, user's verbalisation does not only provide information about what users do, their verbal utterances provide a deeper insight into their task solving

processes, this enables the facilitator to assess the user experience and determine usability problem (Hertzum et al., 2015).

The concurrent think-aloud utility involves different types of instruction. These has been discussed in detail in section 2.5.2.1.

Also, Zhao and McDonald, (2010) compared the classic and a relaxed think-aloud with the aim to explore the impact of think-aloud style on the nature of the utterances produced by participants and the usefulness of those utterances for usability analysis. Findings indicated that the interactive think-aloud led to the production of more utterances than the classic think-aloud, such utterances categories include problem formulation, causal explanation, user experience and recommendation which could be used in usability problem analysis. However, no significant difference was found between interactive and classic think-aloud for utterance categories such as action description, reading and task confusion.
Given the current situation, it is important for usability practitioners to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity.

### 2.8.1 Utterance Categorisation

The category of "utterance" has been a key focus in usability testing research, specifically within the context of 'think-aloud' protocols (Boren & Ramey, 2000). The examination of participant's verbal expressions provides researchers with insights into cognitive processes and problem-solving strategies (Ericsson & Simon, 1980). Many past studies have identified different categories of utterances. For example, Boren & Ramey (2000) divided utterances into problem-related utterances, which identify potential issues with the interface, and non-problem-related utterances, which include comments about aesthetics or general thoughts about the system. These protocols often rely on the analysis of users' spoken thoughts, or utterances, as they interact with a system or perform a task (Nielsen, 1993).

Similarly, Hertzum and Holmegaard (2015) employed three categories in their research: success-related, failure-related, and process-related utterances. Research has shown that different categories of utterances can provide valuable insights into the user's cognitive processes and problem-solving strategies (van den Haak et al., 2003). For example, "description" utterances involve the user describing what they see or do, while "explanation" utterances provide insight into users' underlying decision-making processes (Branch, 2000).

Previous studies within the field of usability, also suggest that critical and emotional utterances, in which users express opinions, frustrations, or satisfaction with the system, can provide particularly valuable insights for usability professionals (Van den Haak et al., 2003; Olmsted-Hawala et al., 2010).

These utterance categorisations allow for an organised and structured analysis of user feedback, and they form the basis for the utterance categories employed in the first and second studies in this thesis.

## 2.9 The Substitution of Explicit Instructions for General Instructions

According to Ericsson & Simon, (1984, p. 80) participants think aloud protocol can be influenced by the exact wording of the think aloud instructions. Ericsson and Simon advocate caution concerning changing the verbalisation instructions in the light of evidence that this may change the structure of the thought process itself. Thus, usability researchers are advised to adopt the general instructions, which only requires participants to think aloud during task performance and a reminder to “keep talking” when participant fall silent with no further prompt.

In an attempt to utilise Ericsson and Simon’s think-aloud guidelines Cotton and Gresty, (2006) encountered problems such as participants’ not knowing what kind of thoughts to

articulate in response to the general instruction as recommended by Ericsson and Simon. They increased the level of guidance give to participants by developing a range of other prompts that might help them collect the type of data they needed.

The prompt used is as follows: “How are you deciding where to go? What do you think of the information in this section?” (p. 49). It is important to know that the prompts were raised as appropriate with each participant and there was no set time schedule to prompt participants’.

McDonald et al., (2013) conducted a study to investigate the relationship between think-aloud instructions, task difficulty and performance. The following explicit instruction was used to request participants thought processes “I want you to think-aloud, explain your link choices as you complete each task”.

Gerjets et al., (2011) conducted a study which compared explicit evaluation instructions and classic thinking-aloud instructions with an expectation that explicit instructions will influence participants’ thought process during the performance of complex search task. The following instruction was used to request participants’ thought process: "mention the evaluation criteria you apply to search results and to assess web pages" (p. 223)

McDonald and Petrie, (2013) conducted a study on the effect of global instructions on think-aloud testing to determine if the classic think-aloud and a think aloud with an explicit instruction will lead to different task solving performance when compared to silent. They instructed participants to express their feelings about a website. The following instruction was used for the explicit condition: "I would like you to think-aloud. I would like you to tell me the things that you like the things that you dislike or finding confusing about the site” (p. 2942).

In addition, Barnum, (2010) on her published book “Usability Testing and Research” replaced the Classic think-aloud instructed as recommended by Ericsson and Simon with an explicit instruction which explicitly request participants to express their emotions and give justification

for their actions. The following explicit instruction was suggested: "We want to know what you expect to happen when you make a choice and whether it meets with your expectations or not. We want to know what surprises, what delights, what confuses or even frustrates you, and why" (Barnum, 2010, p. 205).

Evidence from the above studies indicated that, test facilitators give explicit instruction in accordance with the type of data they anticipated to elicit. The present study will adopt the explicit instruction used my McDonald and Petrie, (2013) because it gives participants a clear view of the extent to which they can used a specified product to achieve specified goals and suggest recommendations where possible, thus giving them the reassurance that the study is a test of the user-friendliness of a product and not a test of their ability to use a specified product.

### 2.10 The Classic Framework of Usability Testing: Its Use and Abuse

Since the first use of think aloud for usability testing, the work of Ericsson and Simon (1993) has been cited by the majority of researchers as the theoretical foundation for using think aloud for usability testing (Boren and Ramey 2000). However, a comparison of the methods and practises used in these evaluations shows inconsistencies between Ericsson and Simon's model and observed practises in human–computer interaction studies (Boren & Ramey, 2000; McDonald, Edwards, & Zhao, 2012; Nrgaard & Hornbk, 2006).

In this section the author will classify the use and abuse of the classic framework of usability testing into four themes: (i) the use and omission of a practice session, (ii) the use of think-aloud demonstration, (iii) think-aloud instructions, (iv) disparity of think-aloud reminders (v) textbooks recommendations on how to implement the think-aloud method during usability testing.

### 2.10.1 The use and omission of a practice session

Some usability studies includes participants think-aloud practice session (Van Kesteren et aI., 2003; Krahmer and Ummelen, 2004; Hertzum et aI., 2009; Karahasanovic et aI., 2009; Cooke, 2010; Olmsted-Hawala et aI., 2010a, b McDonald and Petrie 2013; Zhao and McDonald, 2010; McDonald et al., 2015; Fan et al., 2019), while others did not report this (Grossman et aI., 2009; van den Haak et aI., 2009; Peute etal.,2010; Willis and McDonald, 2016).

### 2.10.2 The use of think-aloud demonstration

The majority of studies made no mention of demonstration, (Hackman and Biers, 1992; Ohnemus and Biers, 1993; van Kesteren et aI., 2003), although, some did include a demo-video to prepare participants for the desired verbalisations (Hackman and Biers, 1992; Ohnemus and Biers, 1993; Lee, Knowles, and Whitehead, 2019). Others recommended the use of video demonstrations in their textbook (Nielsen, 1993; Dumas and Redish, 1999; Dumas and Loring, 2008). Although, Think-aloud demonstrations are not part of Ericsson and Simon's approach, Findings from McDonald, Edwards and Zhao, (2012) indicates that 14 percent of respondent used video-demo.

### 2.10.3 Think-aloud instructions

The majority of usability studies make use of a general instruction to asked participants to think-aloud, (i.e. van den Haak, de Jong and Schellens, 2007; Hertzum et aI. 2009), some of them used a more specific set of instructions. (Shrimpton-Smith, Zaman and Geerts, 2008; Wright and Converse, 1992). While others, did not report the type of instruction used in their studies (Held and Biers, 1992; Ohnemus and Biers, 1993; van Kesteren et aI., 2003; Als et al, 2005a, b; Edwards and Benedyk, 2007; Jensen, 2007; Cooke, 2010).

### 2.10.4 Disparity of think-aloud reminders

Concerning the use of think-aloud reminders, some studies used a simple "keep talking" reminder only (Krahmer and Ummelen, 2004; Hertzum et al., 2009; Peute et al., 2010; ADD SHARON CITATIONS ), while Wright and Converse (1992) used "what are you thinking about?" Also, some studies makes use of interventions which includes offering help (Karanasanovic et al., 2009; Grossman et al., 2009) and probing for further explanations (Wright and Converse, 1992). Other failed to report their use (Held and Biers, 1992; Eger et al., 2007; van den Haak, 2008).

### 2.10.5 Evidence from research has shown methodological irregularities

The literature shows that not only does the implementation of the concurrent think aloud procedure vary, but the amount of detail offered about the process implementation is also inadequate. Due to omitted information, it was difficult to determine whether or not the research conducted certain activities that were not published.

### 2.10.6 Textbook Recommendations on how to use the think-aloud method.

The variety of advice on how to apply the think-aloud methods provided in the many texts may impact on the way it is being implemented in practice. For instance, Dumas and Redish (1999) suggested the use of a general instruction which contrasts with Ericsson and Simon's established guidelines. Similarly, Barnum, (1999) recommends that test facilitator to give explicit instructions to participants, such as instructions that direct participants to verbalise their likes, dislikes and emotional response during verbalisation. Also, some books recommended a think-aloud demonstration (Nielsen,1993; Dumas and Redish, 1999; Dumas and B. A. Loring, 2008). Whereas other do not (Barnum,2020). Some recommended that test facilitator to conduct a think-aloud practice session with participants (Barnum, 2020; Dumas and Redish, 1999; Dumas and B. A. Loring, 2008). Also, some textbooks recommended the

use of probes, prompt and interventions which is direct violation of Ericsson and Simon's established guidelines (Ericsson and Simon, 1993).

### 2.10.7 Methodology and Evaluators Effects

Researchers have examined the difference between strict and relaxed TA, where "relaxed" refers to loosening up the procedures of Ericsson and Simon's strict TA protocols, with resulting differences in participant descriptions of how to do TA, practise times, reminder types, prompting intervals and intervention styles (Boren & Ramey, 2000). For instance, some moderator gave detailed and explicit instructions on how to think aloud, whereas others only gave simple instructions. And moderators used neutral affirmations such as "Yeah...," but also positive affirmations such as "Great!" (Molich, et al., 2020).

The moderator's expertise and abilities have a major impact on the outcomes of usability research (Barnum, 2011; Dumas & Loring, 2008; Dumas & Redish, 1999). For instance, moderators can directly influence outcomes by asking biased questions and offering premature assistance, and cognitive biases such as the confirmation bias can have a subtle influence on verbal behaviours and task performance. Twelve think-aloud usability sessions were examined by Hertzum and Kristoffersen, (2018) to determine the type of verbalisations moderators made before, and during think-aloud by classifying the moderator's comments from 12 test sessions. Findings indicated that the most frequent moderator verbalisations during test tasks were affirmations (38%) followed by task instructions (32%) and cues for reflection (16%). The moderators spoke less throughout the tasks than they did before and after the test session with the most common verbalisations being "Mm hm," "Okay," and "Uh-huh" accompanied by instructions and prompts for reflection. Overall moderators, verbalised more than usual.

The usability of the same website, Microsoft Hotmail, was assessed by nine different organisations. The findings show a wide range of differences in methodology and execution, and problems identified. 310 separate usability issues were identified by the organisations.

Just two issues were reported by six or more organisations, while 232 issues (75%) were reported uniquely, meaning no two teams reported the same issue. Some of the unique findings were deemed serious. More importantly the tasks used by any or all departments yielded wildly varying outcomes – about 70% of each task's findings were unique. (Molich et al., 2007)

### 2.11 Think-aloud and Tasks

In terms of think-aloud verbalisation and task, Davis and Bistodeau, (1993) stated that, many thought processes in working memory are not verbalised, either because they are unconscious, such as identification of common words and images or because their "intermediate" processing occurs so rapidly that there is little time to verbalise it. Hence, practitioners should carefully choose they study tasks. Nonetheless, before designing a study which involves think-aloud methods, practitioners need to decide on the task-type and level of difficulty (Charters, 2003). Ericsson and Simon (1980) found that demanding tasks trigger a “high cognitive load” conflict with verbalisation because other processes crowd verbal information out of working memory. On the other hand a simple task may also be unsuitable as “the closer readers’ activities come to automaticity, the more difficult it may be for readers to explain these automatic or near-automatic happenings. (Pressley & Afflerbach, 1995, p. 132).

Hertzum et al. (2009) proposed three alternatives. Firstly, to stick with the classic solution since it provides more certainty in terms of validity and reliability. Secondly, to accept that the traditional approach is irrelevant for usability research and instead focus our efforts on understanding and improving the relaxed think-aloud method. The last option would be to abandon the use of think-aloud entirely.

Medina (2008) also suggested that tasks be broken down into smaller units such that they can be worked on one at a time to avoid overloading working memory, and that tasks should be

written on a piece of paper that can be conveniently referred to in order to free up space in working memory so that higher-level thought can occur.

Another area of investigation has focused on the discrepancies in verbal data obtained in moderated and unmoderated usability studies research that has yielded contradictory results (Lewis, 2012, 2014).

## 2.12 Concurrent Think-Aloud in Usability Testing: What the Future Holds

Over the last few decades, user experience and usability practitioners has undergone changes due to the following: The implementation of unmoderated usability testing methods (Albert, Tullis, & Tedesco, 2010). The adoption and use of agile and lean methodologies in some development settings. Methodologies that can make it difficult to incorporate user experience input and assessments (Stellman & Greene, 2014). The extension of user researchers' concerns beyond the domain of traditional usability to broader concepts of user experience (Diefenbach, Kolb, & Hassenzahl, 2014).

Evidence for literature has shown that the concurrent think-aloud method is the most used techniques by practitioners when conducting a usability testing (McDonald et al., 2012) and many UX practitioners' current practices deviate from Ericsson and Simon's three guidelines. Based on findings from literature the author will consider different themes such as: tasks; methodology and evaluators effects.

## 2.14 Summary

In accordance with the literature review, a thorough examination of the concurrent think-aloud and its procedural factors and issues surrounding reactivity is required. The outcomes ought to shed light on how the concurrent think-aloud procedure should be implemented when conducting usability testing.

A suitable study strategy must be chosen, and numerous evaluation metrics must be constructed, in order to examine the impact of concurrent think-aloud and its methodological alterations on usability evaluation. The next chapter discusses the approach for this research and explains the primary criteria utilised in this think-aloud study to evaluate the concurrent think-aloud techniques and the reactivity concern.

Also, the authors of this thesis have performed research and usability studies that contributes to our current understanding of the effects of these evolutionary pressures and points the way to effective methodological adaptations that will benefit both researchers and practitioners. See details in originality in chapter 7 section 7.5.

# CHAPTER THREE

## 3.0 METHODOLOGY

### 3.1 Overview: Methodological Exploration Based Upon Past Studies

This chapter details the methodological approach used to obtain empirical data, the rationale for considering each experimental design, and the approach used for analysing the qualitative data were discussed in section 3.4.1. The adopted methodology for each study varies with respect to the study requirements, research questions and hypothesis. Hence, each methodology was reported in the relevant chapters.

### 3.2 Introduction

This thesis investigates the impact of task-types on the utility of the think-aloud method within usability testing. The aim of the research and its underlying research question indicated that an experimental approach should be adopted. For this thesis, there are three rationales for choosing an experimental approach.

First, this research involves a study that compares different variants of the concurrent think-aloud method with different procedural medications (e.g., explicit instructions) using different task types to examine their relationship and impact on the resultant data. For example, study one will address RQ1, i, and to partially address ii by comparing CTA and a silent condition with fact-finding tasks, and CTA and silent with sensemaking tasks, the reason being to investigate the impact of different task type on reactivity in CTA. See section 6.2 for details. To fully address RQ ii, a second study (study two) will compare CTA and a silent condition with fact-finding tasks; CTA and silent with search tasks and CTA and silent with sensemaking tasks. The Independent variables: Classic Concurrent Think-aloud and Silent Working.

Secondly, the think-aloud method is mostly used in usability laboratory settings, (Symonds, 2011). Hence, it is appropriate to adopt the use of an experimental approach for the data collection. Also, the research questions that the studies intend to answer requires the use of

large-scale qualitative analysis of the elicited utterances obtained from the usability test that will be conducted.

Finally, researchers have applied different measures to assess the impact of the think-aloud method within usability testing, however, none has investigated the impact of task-types on think-aloud. In general, different research has different aims and objectives which sometimes contribute to the limited knowledge on the issue of think-aloud and reactivity.

The subsequent section discusses various experimental design issues relevant to think-aloud studies. Detailed experimental design information is documented separately, each in its empirical study chapter, the rationale for doing this is because each study uniquely addresses specific research questions that are relevant to the study, hence, the experimental design and the measures that are examined varies with respect to the study's research questions.

## 3.3 Research Philosophies

### 3.3.1 Experiment Variables and Designs

An experiment in psychology involves establishing independent variables, measuring dependent variables and controlling extraneous variables.

According to Kirk, (2012) an experiment is an investigation in which a hypothesis is scientifically tested, an independent variable (the cause) is manipulated, and the dependent variable (the effect) is measured, and any extraneous variables are controlled.

#### 3.3.1.1 Independent Variables

Independent variables refer to the creation of experimental conditions or comparisons that are under the direct control of the researcher. Manipulated independent variables can involve participants placement in different conditions, assigning them different tasks, or giving them different instructions. For example, McDonald, McGarry, and Willis, (2013) conducted a study

to investigate If an explicit explanation-based think-aloud instruction leads to differences in navigation performance over the classic think-aloud method. The independent variables (i) Tasks success which encompasses high scent and low scent tasks, (ii) Time on task.

The experimental conditions or comparisons that are used in studies within this PhD thesis are procedural methodological changes to the classic think-aloud.

#### 3.3.1.2 Dependent Variables

Dependent Variables are what the researcher intends to measure within the experiment, and they are affected by the independent variables, such variables can be time on task, mouse click counts, or the number of abandoned tasks. For example, McDonald, McGarry, and Willis, (2013) focuses on think-aloud instructions, task difficulty and performance. These are important measures that they thought could be affected by the independent variables. In this thesis, they were the three themes mentioned in section 3.5.2: Performance, subjective testing experience, participants' verbalisations. These specific measures however varied among the empirical studies conducted within this thesis, thus, they are covered in relevant chapters.

#### 3.3.1.3 Extraneous Variables

These are any variables that are not of interest to the researcher, but which might influence the behaviour being studied if they are not controlled properly (Danziger and Avnaim-Pesso, 2011). As long as these are held constant, they present no danger to the study. However, if a researcher fails to control extraneous variables, they can then influence the behaviour that is being tested or measured in some systematic way. The outcome is called confounding. A confound refers to any uncontrolled extraneous[1] variable that ‘‘covaries’’ with the independent variable and could provide an alternative explanation to the outcome of an experiment. That

[1] Extraneous or confounding variables are any other variable that could affect the dependent variable but is not explicitly included in the experiment. Therefore, all the other variables that could affect the dependent variable to change must be controlled.

is, a confounding variable changes simultaneously with the independent variable. (i.e., they ‘‘covary’’) and, consequently, its effect cannot be separated from the effect of the independent variable.

Therefore, when a study has a confound, the results could be due to the effects of either the confounding variable or the independent variable, or some combination of the two, and there is no way to decide among these alternatives, this means confounded studies are uninterpretable. Since dependent variables are the behaviours that are measured in a study, they must be defined precisely to avoid confound.

According to Kardes and Herr, (2019) to make an experiment valid and less biased, it must be objective. The opinion and views of the researcher should not be considered relative to the result of a study. A systematic and carefully planned process needs to be applied to deliver reliable, valid and repeatable experimental results. Also, it is important to anticipate the type of data that is to be collected to ensure the most appropriate method is applied in relation to the research questions and the number of participants that is enough to obtain a reliable and valid result. In this context, to detect reactivity if there is reactivity to be detected.

### 3.3.2 Mixed Design: Qualitative and Quantitative

Mixed methods have been described as the “third methodological movement” following quantitatively and qualitatively oriented approaches (Teddlie and Tashakkori, 2003). According to Creswell et al., (2003) a mixed methods study involves the collection or analysis of both quantitative and/or qualitative data in a single study in which the data are collected concurrently or sequentially, are given a priority, and involve the integration of the data at one or more stages in the process of research.

In this section, we will examine mixed methods research design, the advantages it provides, and its implementation in diverse fields, including the domain of human-computer interaction.

There are different types of mixed methods research designs, such as sequential explanatory, sequential exploratory, concurrent triangulation, and concurrent nested designs. In the sequential explanatory design, researchers gather and analyse qualitative data before quantitative data. Conversely, in the sequential exploratory design, quantitative data is collected first, followed by qualitative data. In a triangulation design, both types of data are gathered at the same time and analysed separately. Integration between the two types of data happens during interpretation. Finally, in nested designs, one type of data (qualitative or quantitative) is dominant and is supplemented with the other data type. When resources are constrained and mixed methods are preferred, nested designs can be used in bigger studies. These designs allow for adequate levels of within-subject analysis of both qualitative and quantitative data (Yoshikawa, Weisner, Kalil, & Way, 2008).

Mixed methods research offers numerous benefits over single-method approaches. The integration of qualitative and quantitative data allows researchers to gain a more comprehensive understanding of complex research questions. The qualitative data can provide rich contextual information, which helps to explain the quantitative data. Additionally, the use of both methods can enhance the validity and reliability of research findings by triangulating the data. Triangulation involves the use of multiple sources of data to confirm or refute research findings, leading to more robust and trustworthy research outcomes.

Mixed methods research has been widely used in various fields, including human computer interaction, healthcare, social sciences, education, and psychology.

In psychology and human computer interaction, mixed methods have been used to explore participants behaviour when using different think-aloud variant or to test for reactivity. For example, a study by Bowers and Snyder (1990) used a mixed factorial design, using a between-subject treatment of verbal protocol, because of the likelihood of unequal transfer of

training between the verbal protocol conditions. Findings indicated that, due to the level of task difficulty, participants in the concurrent condition were forced to give verbalisations that requires further cognitive processing, thus inducing reactivity. Similarly, Eger et al., (2007) used a nested design with mixed within and between-participant factors. Each participant produced a think-aloud protocol with one search engine and produce one of the two types of retrospective protocols with the other search engine. Their findings indicated that, the eye-cued method identified more usability problems than the think-aloud or screen-cued methods and fewer participants completed the search task on the Think-aloud condition, indicating the reactivity of the technique. Also, McDonald, McGarry and Willis (2013) adopted a mixed factorial design, a between subjects variable of CTA vs Explicit think-aloud and a within subject variable of high scent vs low scent. Findings indicated that, for the low difficult tasks, there was no difference in task success however, for complex tasks there were differences, as participants in the classic condition completed fewer tasks successfully and engaged in more link traversals.

In social sciences, mixed methods have been used to explore various research questions, including social inequality, social movements, and social networks. For example, a study by Kelleher and colleagues (2013) used mixed methods to explore the experiences of mothers with postpartum depression. The study found that the stigma associated with depression prevented many mothers from seeking help, and that social support was critical in facilitating recovery. In education, mixed methods research has been used to explore teaching and learning processes, student experiences, and educational outcomes. For example, a study by Creswell and Plano Clark (2011) used mixed methods to explore the impact of teacher professional development on student achievement. The study found that the professional development program improved teacher knowledge and skills, which led to improved student achievement.

A mixed method approach can offer a more comprehensive understanding of complex research questions and can also enhance the validity and reliability of research findings. (Brannen, 2005).

#### 3.3.2.1 Role of Experimental Design

To conduct an empirical study, different approaches can be used, although it is important to apply to the most appropriate method concerning the research questions for the underlying study. These approaches are between subjects (independent measures) and within-subjects (repeated measures).

#### 3.3.2.1 Within-Subjects Designs (Repeated Measure)

In within-subjects designs, all participants are exposed to all experimental conditions. In essence, each participant serves as his or her control (Greenwald, 1976; Shani and Gunawardana, 2011).

***Using Within-Subjects Designs***

The characteristics of within-subjects designs are:

- Each participant is tested under each experimental condition.
- Therefore, the scores in each condition are correlated with the scores in each other condition.
- The critical comparison is the difference between correlated groups on the dependent variable.

In within-subject designs, a single sample of participants is exposed to all of the conditions of the experiment. Because the same participants are in all conditions, the experience each participant has in one condition might affect how that participant responds in subsequent conditions. Thus, if differences between the conditions are found, they might not be due to the

independent variable manipulation, but to the confounding effects of one condition on later conditions. These confounding effects are called sequencing effects, and they must be controlled. One of the controls for sequencing effects is counterbalancing. In complete counterbalancing, the order of presentation of conditions to participants is systematically varied so that: (1) each participant is exposed to all of the conditions of the experiment, (2) each condition is presented an equal number of times, (3) each condition is presented an equal number of times in each position, and (4) each condition precedes and follows each other condition an equal number of times.

***Analysing within-subjects design***

The most appropriate statistical analysis for a single-variable, within-subjects experiment is a repeated-measures ANOVA, which takes into account the fact that the measures are correlated. The advantage of a within-subjects design is that it effectively equates the participants in the different conditions before the experimental manipulation. Therefore, the single largest contributing factor to error variance—individual differences—has been eliminated. Reducing the error variance increases the F-ratio. (Since the individual difference portion of the error term has been removed, the denominator in the F-ratio will be smaller and, therefore, the F will be larger.) This means that the procedure will be more sensitive to small differences between the groups (Brauer and Curtin, 2018).

In the repeated-measures ANOVA, the between-groups and total sums of squares are computed in the same way as in a simple one-way ANOVA. However, the within-groups sum of squares is split into two terms: a subject's term (the individual differences component of the within-groups variability), and an error term (what is left of the within-groups variability after the individual differences component is removed).

***Strengths of Within-Subjects Design***

There are several advantages of within-subjects designs:

i. Because the same participants are in each condition, the groups of participants are equivalent at the start of the study.
ii. Within-subjects designs are more sensitive than between-subject designs to effects of the independent variable manipulation.
iii. Within-subjects designs are more efficient than between-subjects designs in two ways: (a) fewer participants are needed, and (b) instructions often need be given only once per participant, instead of once per participant for each condition.

***Weaknesses of Within-Subjects Design***

There are several disadvantages of within-subjects designs, all of which stem from the fact that each participant is exposed to each condition. Participants' experiences in one condition might affect their responses in one or more of the subsequent conditions. Therefore, differences between groups might not be due to the independent variable manipulation but, rather, to the confounding factor of sequencing effects, which include practice effects and carry-over effects. Practice effects are caused by the participants' practice and growing experience as they move through the sequence of conditions. This effect is due to the participants' growing general familiarity with the procedures. These may be positive or negative practice effects. Carry-over effects are sequencing effects due to the influence of a particular condition or combination of conditions on responses to the very next condition.

There are two general types of controls for sequencing effects: (1) holding the extraneous variable constant, and (2) varying the order of presentation of the conditions (counterbalancing or random ordering). Counterbalancing can be complete or partial. Latin square designs are examples of partial counterbalancing. Trials can also be randomized within blocks. However, if strong carry-over effects are expected, a within-subjects design is not recommended, even if the above controls are included.

#### 3.3.2.2 Independent or Between-subjects design

The participants are assigned into equal-sized groups and each group receives only one condition (different treatment or no treatment) this prevents condition cross-contamination between the groups.

***Using Independent or Between-subjects design***

The general goal of the between-subjects experiment is to determine whether differences exist between two or more treatment conditions (e.g., an author may want to compare two teaching methods (two treatments) to determine whether one is more effective than the other)

***Strength of Between-subjects design***

- Allows a researcher to look at the effects of treatment in isolation.

***Weaknesses of Between-subjects design***

- Requires a large number of participants which depends on the study's methodology.
- It is difficult to endure that the groups are equivalent.

#### 3.3.2.3 A Factorial Design

Used where there are several independent variables, and the researcher is interested in their combined effect on the dependent variable. A study that involves only one independent variable is called a single-factor design. A study with more than one independent variable is called a factorial design.

***Strength of Factorial design***

- Greater precision can be obtained in estimating the overall main factor effects.
- Make research cheaper
- Allow many levels of analysis
- Highlights the relationships between variables

- Allows the effects of manipulating a single variable to be isolated and analysed singly
- Interaction between different factors can be explored.
- Additional factors can help to extend the validity of conclusions derived.

***Weaknesses of Factorial design***

The main disadvantage is the difficulty of experimenting with more than two factors or many levels. A factorial design has to be planned meticulously, as an error in one of the levels, or the general operationalisation, will jeopardize a great amount of work.

Other than these slight detractions, a factorial design is a mainstay of many scientific disciplines, delivering great results in the field.

#### 3.3.2.4 Matched-Subjects Design

Matched-subjects designs use different participants in each group, but the participants have been closely matched before assignment to conditions.

The characteristics are:

i. Each participant is exposed to only one level of the independent variable.
ii. Each participant has a matched participant in each of the other conditions, so the groups are correlated.
iii. Only one measurement per participant on the dependent variable is used, and the analysis takes into account the matching.
iv. The critical comparison is the difference between the correlated groups.

***Using Matched-Subjects Designs***

A matched-subjects design is used when the author wants to take advantage of the greater sensitivity to independent variable manipulations, but cannot, or chooses not to, use the within-subjects design. The most common example is when the manipulations would cause severe sequencing effects.

Participants should be matched on relevant variables. A variable is relevant if it can affect the dependent variable in a study. That is, the important variables on which to match are variables that are strongly related to performance on the dependent measures. Matching on more than one variable can become difficult.

***Analysing Matched-Subjects Designs***

In analysing results of matched-subjects designs, it is necessary to maintain the ordering of the data (i.e., who each participant is matched with). The same statistical procedures used for within-subjects designs are appropriate for the matched-subjects designs.

***Strengths and weaknesses of a matched-subjects design***

Within-subjects and matched-subjects designs have similar strengths, but different weaknesses. Both have good sensitivity to small differences between conditions because the groups are equivalent (or even identical). Therefore, smaller numbers of participants are needed. An advantage of the matched-subjects design over the within-subjects design is that no problems are resulting from practice and carry-over effects. Therefore, procedures such as counterbalancing are not needed. The most obvious disadvantage of the matched-subjects design is that it requires a good deal of effort to match participants. Also, the requirements of matching might eliminate many potential participants because suitable matches cannot be found for them. In such cases, we may be better off using a large group of randomly assigned participants in a between-subjects design.

### 3.3.3 Usability Evaluation in Sunderland University

The “gold standard” in usability testing is traditional lab-based usability testing (Landauer, 1995; Newman, 1998). The usability evaluation process in University of Sunderland follow much the very process in which a basic usability test is conducted but in a more rigorous and robust way compare to what is been done in the UX industry. The process involves different

phases: planning; testing; analyse and communicate phases. These phases are further broken down into small units which helps to standardise the process so that the results are consistent and reliable. The available resources is also put into consideration, the cost of the research and the time estimate.

### 3.3.4 The Planning Phase

Our evaluation process at the University of Sunderland, planning the details of a usability test season is the most crucial part of the entire process because the decisions that are made at the start of the testing process will determine the outcomes of the usability test. The planning phase involves setting goals and scope of the research. We start with problem definition, the purpose of the research study and then formulate research questions and hypothesis. Then we get familiar with the tested product through usability inspections to get some early notions on usability issues. The information gathered will enable us to decide on what method will be suitable to answer the underlying research questions. The overall process is presented in figure 1.

Our evaluation process focusses on three core elements of a usability test: tasks; participants and the test facilitator. This is because a usability test is a task-base process which involves a participant and a facilitator who assigns tasks and observers the participant's behaviour. Our evaluation process takes participants recruitment crucial as it involves recruiting representative participants of the product or service that is being studied. The lab set up is simple with an eye tracker which is embedded on the computer (Tobi) which is barely noticed by the participants. The metrics for each study depends on the research questions and this is further considered when formulating the tasks for each study.

The planning phase is usually finalised with documentation and materials for the study, these includes participants' information sheet; consent form; pre-test questionnaire; screening form; user profile form; TLX forms; after scenario questionnaire etc.

### 3.3.5 The Test Phases

The testing phase starts with a pilot test. Usually, a minimum of three pilot test is conducted this is to validate the study tasks and the test documents and the overall test process meets expectations. When it comes to conducting the usability testing session, we ensure facilitators stick to a strict protocol with each participant. This protocol allows for some personalisation while ensuring that each participant has a consistent experience.

Both in the pilot study and the actual test, we introduce a warmup session to make sure participant is physically comfortable with the testing setup (chair, desk height, mouse placement, etc.) and that they understand what will take place during the session. we ask them some friendly conversational questions, such as how far they've travelled to get to the lab, whether they've previously done usability test, and so on. Then we collect pre-testing data, before transition into the first task we give them a set of instructions, to think-aloud and let them know if there will be a post-task questionnaire and a post-test questionnaire or a very short interview at the end of the test session. see more details in chapter 4, section 4.4.2, also, see details of test procedures for the first study in section 4.4.5 and for the second study in section 5.4.5.

### 3.3.6 The Analysis and Report Communication Phases

Finally, after data collection, we try to analyse the data as soon as possible after testing so that the observations are fresh in mind regardless of the recordings. First, we review the original research aim and then identifies areas of interest, transcribe participant data.

We organise the data, draw conclusion, prioritise the issues and compile a report. see details in section 4.6. The report is communicated via academic journals, presentation on slide deck or video clips and sometimes workshop on improvement.

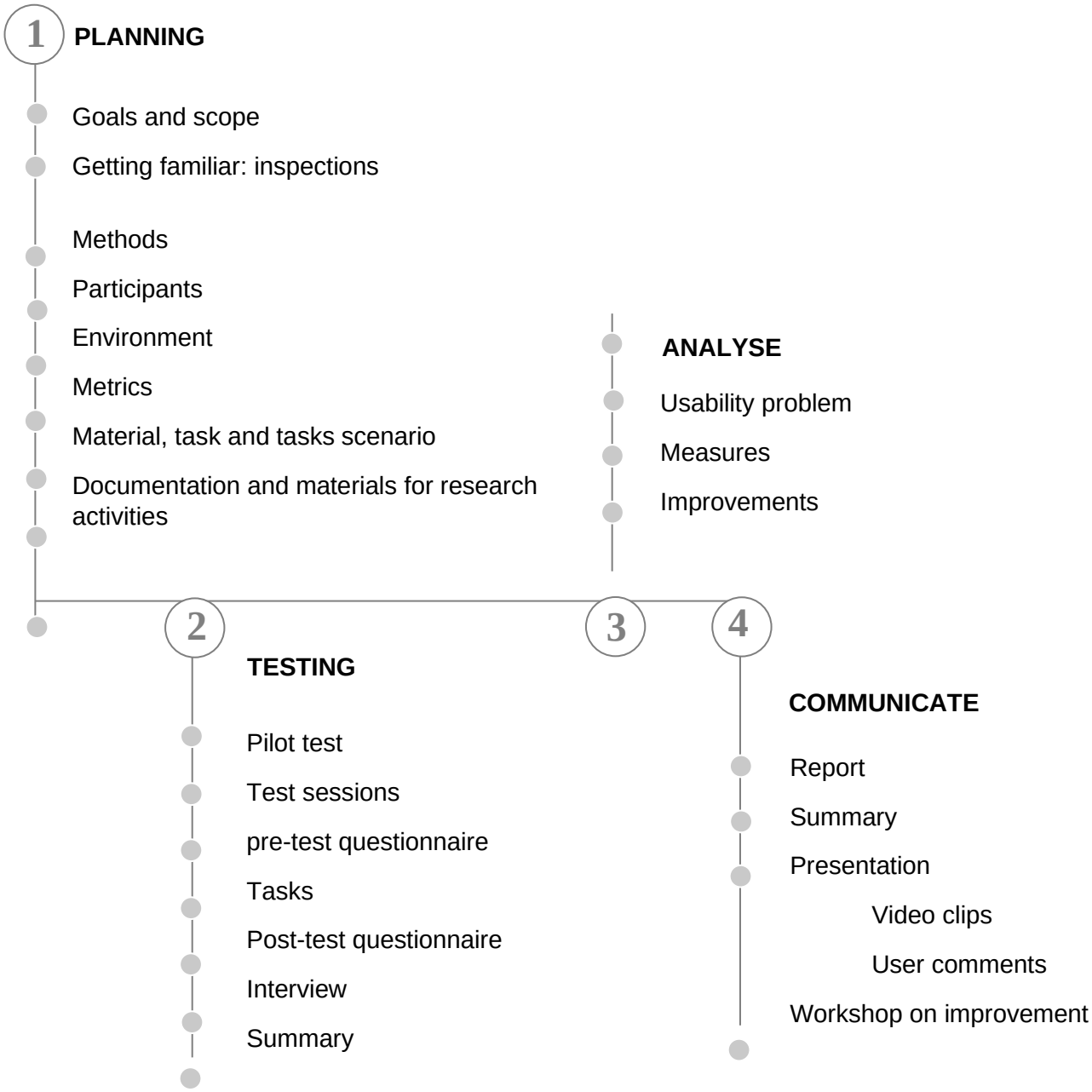



*Figure 1:Usability evaluation process in University of Sunderland*

### 3.3.7 Data Capture: Screen recording and Interviews

This research will make use of screen recording as part of the method to capture data from participants during usability testing sessions. It is crucial to gather precise usability test data, communicate unbiased results, and provide well-supported recommendations when identifying usability issues during a usability test or for website modifications (Goodwin, 2005).

Data capture during usability testing is a crucial part of evaluating user experiences. screen recordings, after scenario Interviews, and other usability metrics are commonly used techniques for capturing data during usability testing (Boren and Ramey, 2000). By recording the user's screen, researchers can gain insight into how users interact with a product or system. Screen recordings allow researchers to observe users' behaviour and identify usability issues in real-time (Zhao and McDonald, 2010).

During the study, it is necessary for the researcher to record a number of things using screen capture software, video and audio. This recorded data will be stored securely, and can only be accessed by the researcher. See further details in study information sheet about consent and ethical considerations in Appendix, page 219 and 235 respectively.

The participant sessions were recorded (video and audio) including the screen using TechSmith Morae on the Tobi Studio Eye tracker machine. These methods were used in this research as part of the data capturing process for study one and two. See details of study procedure for study one and two on section 4.4.5 and 5.4.6 respectively.

Interviews are a useful way to gain insight into users' experiences during usability testing. Semi-structured interviews were also conducted to obtain qualitative data relating to participants' experience with the two think-aloud styles. Interview questions were carefully phrased and piloted, to ensure the wording of questions would not introduce any potential biases.

### 3.3.8 TLX: Instrument Which Supports Data Capture.

The NASA Task Load Index (TLX) is a widely used, subjective, multidimensional assessment tool that measures perceived workload in order to assess a human's interaction with various systems. Developed by Sandra G. Hart and Lowell E. Staveland in 1988, it has been extensively used in research for evaluating different aspects of workload associated with a task or a system.

The TLX gives an overview of workload by considering six sub-scales:

1. Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?

2. Temporal Demand: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?

3. Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter or yourself? How satisfied were you with your performance in accomplishing these goals?

4. Effort: How hard did you have to work (mentally and physically) to achieve your level of performance?

5. Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

In this thesis, the TLX was used to gain insights into perceived workload and how it may affect performance and user experience in the studies conducted within this thesis. It has been employed across a wide range of fields, from aviation and healthcare to human-computer interaction and driver behaviour studies. This tool has proven especially useful in ergonomics and design, helping to optimise the balance between system demands and human capacities. See details in appendix page 230 and 246 respectively.

It's important to note that TLX measures perceived workload, not the workload itself. Therefore, the results can vary depending on the individuals involved and their perception of the task's difficulty.

### 3.3.8 Mode of Piloting

In order to conduct a successful usability test, the formulated task set need to be piloted to identify possible misleading tasks and to ensure simplicity in terms of participants understanding of the tasks. The tasks were designed to avoid answers overlapping from earlier to later tasks. In terms of the explicit think-aloud instruction, a short instruction was written on every task to ensure participants do not forget the essence of the task performance.

The tasks and test procedure were piloted with three participants prior to the commencement of data collection for study one. The test lasted for 30 minutes, including informing participants about the purpose of the test and ethical consideration to ensure participant overall safety no test forms were filled and a debrief section about the test.

For the second study, a first pilot test was conducted with one participant, the test lasted for 40 minutes, although this includes a debrief section about the test and ethical consideration to ensure participant overall safety no test forms were filled. Feedback received including the use of simple language to enable users understand the tasks, research supervisor strongly recommends another pilot test which should include participants debriefing and filling of all corresponding forms that will be used for the actual test to ensure a successful experiment.

A second pilot test was conducted with a different participant with a full test session which include: introduction and debriefing about the Study and ethical consideration of the Study; a practice session for think-aloud condition; tasks performance and post-test questionnaire. Participant filled all required forms, after tasks questionnaire and after test

questionnaire, the pilot study lasted for 66 minutes with minor suggestion to tasks sentences. See table 15 for details pilot test session.

| Test phase | Duration in minutes |
|---|---|
| Introduction and debriefing | 5 - 10 |
| A practice session for Think-aloud condition | 3 |
| Tasks performance | 30 - 45 |
| Post-test questionnaire | 5 – 8 |
| **Total time spend on test** | **43 - 66** |

*Table 3: Pilot test session*

The mode of piloting for study three is documented in chapter six of this thesis, see section 6.4.2.2 for details on running a pilot interview section.

## 3.4 Sampling Approaches: Qualitative Data Analysis

### 3.4.1 Qualitative Data Analysis as Recommended by Chi (1997)

The individual test sessions were transcribed and segmented into individual utterances for fact tasks and sensemaking tasks, the verbal utterances differ in length, however, each comprises of a single theme. The individual utterances were annotated with the participant number and task number.

Simon and Ericsson (1993) opine that context-free coding should be adopted during verbal data analysis in order to limit the likelihood of analyst-induced bias. They argued that, in bringing the raw data to the final categorisation scheme, five criteria should be used to protect the integrity of both the data and the processes it represents, see details in table 1.

However, Yang (2003) argued that Ericsson and Simon's five criteria for encoding verbal data for analysis were not entirely correct when applied to the ill-structured domain.

The subsequent section explains Chi qualitative data analysis and how it is been applied in this research.

### 3.4.2 To Reduce the Sampled protocols

Qualitative data exists in form of narrative text, commonly gathered from interviews, survey questions and recorded observations among other sources, thus the data analysis process can often become complex, laborious and time-consuming as such many authors choose to code only a sample of the data. Authors can choose sampling method depending on their concerns in the research design.

They are three general methods for data reduction which is been used are: (i) Random sampling: a sample is chosen randomly so that each possible sample has the same probability of being chosen. (ii) Systematic sampling: selection of samples according to some non-content criteria such as pauses, changes in speaker, changes in activity. It is also important to note that non-content criteria sometimes do require having a cursory idea of the content with respect to the case of identifying the activity. (iii) Preliminary coding of the entire content set and then more detailed coding on a selected subset.

With regards to this research, this first step: sampling the protocol was not adopted. All participants video data from the think-aloud session was reviewed, transcribed and analysed as the main purpose is to explore patterns of cognitive engagement of every participant, also one cannot justify if a randomly selected sample could be a true representation of entire sample size.

### 3.4.3 To Segment the reduced or sampled protocols (optional)

This involves segmenting the unit of analysis, with regards to this research, where each transcribed video can be broken down into segments which will be further coded into defined categories.

Chi (1997) recommends the necessity to consider the granularity of the segmentation with respect to the size of the unit for analysis such as a sentence, an idea, a reasoning chain or an episode. A coarser segmentation requires less effort; however, there is a tradeoff as to the amount of information that can be obtained. While a fine-grained level segmentation demands significantly more time and effort, however, proved to be more beneficial in terms of producing a more sensitive data

Thus, with regards to this research, fine-grain segmentation was adopted in order to understand participant's utterances and to accurately interpret them so as to obtain the different patterns of cognitive engagement of every participant and to draw a more robust conclusion regarding this research studies.

### 3.4.4 To Develop or choosing a coding scheme

In this phase, the initial coding scheme needs to be developed based on the prior literature and empirical data or established by the author himself, however, as the case may be, the coding scheme can be modified during the process of coding (Chi, 1997). For instance, asking a conceptual question with regards to the research theoretical orientation, the hypotheses, tasks and problem area or content domain could be identified as one of the categories as it has been found from previous empirical studies in investigating participant's cognitive engagement (Chin and brown 2000).

Cooke (2010) suggested five categories: Explanation, Observation, Procedures, Reading and Others. However, these set of categories are generic, hence suggest that a set of categories

may not always represent all verbal utterances but may be appropriate within some context. If authors identify a category that has not been stated from prior studies on cognitive engagement, then this category could be identified as data derived. New categories can be added during the process of coding all empirical data.

With reference to previous research, coding scheme used was too generic and therefore were not directly adopted with this current research.

### 3.4.5 To Operationalise Evidence In The Coded Protocols That Constitutes A Mapping To Some Chosen Formalism

In this phase, every unit of analysis is coded into a defined category indicating the evidence that the coded utterance belongs to a specific category. Therefore, the author is expected to analyse what evidence contained in an utterance can be assigned a specific code (Chi, 2017). Chi highlighted that some of the utterances can be ambiguous, hence, the coders need to determine the extent to which they consider the nearby utterances. She recommended two approaches: Firstly, to consider some of the segmentation that occurs before and after the segmentation that is been analysed, that is making optimum use of the context to improve their understanding of the verbal utterances; secondly, by using only the nearby utterances that surrounds the segment that is been analysed for appropriate interpretation. However, these recommendations contrast Ericsson and Simon's context-free approach.

This research use context-appreciative encoding, hence, factors such as test situation which encompasses: the test session; tasks and participants are required to flawlessly code the verbal utterances in the best possible way in order to accurately represent participant’s cognitive engagement.

It is important to highlight that, the objective of the analysis is not to code the correctness or incorrectness of the represented knowledge but to code the content of the verbal utterance

based on what that utterance indicates about how participants engaged in the task solving process.

### 3.4.6 Depicting the Mapped Formalism (Optional)

This phase is to use a way to depict coded data, which depends on the formalism that has been chosen. Authors can provide a simple table presenting the frequency, time, and length of the identified categories to depict the coded data.

### 3.4.7 Seeking Pattern(S) In the Mapped Formalism

This phase is to seek patterns in the depicted data. For instance, authors can analyse the relationships between various coded categories or draw a model based on the patterns sought by the depicted data.

### 3.4.8 Interpreting the Pattern(s)

In this phase, hypotheses are being tested. Frequency distributions of different coding categories can be analysed using non-parametric methods to explore relationships among categories of responses within samples or across samples. Authors need to select its types of statistical analysis based on their proposed research questions.

### 3.4.9 Repeating the Whole Process, Perhaps Coding At A Different Grain Size (Optional)

Authors can choose to re-code the data at a different grain size or if they want to address different research questions. Since verbal data contain rich sources, new research questions may occur during the process of recoding the data.

With regards to this research, this phase was not adopted as the author assume that recoding the data is unnecessary. It is important to note that, whilst each procedure was introduced separately, the process may be applied in an integral way, for instance, the procedures of segmenting the reduced or sampled protocols, developing or choosing a coding scheme and seeking pattern(s) can be employed together during the coding process.

This research intends to follow the stages of verbal data analysis from the segmentation stage through to the interpretation stage as set out by Chi, Chi (1997). The rationale behind this decision is due to the fact that context-free coding: excluding the problem domain with respect to situation, tasks and users, that is proposed by Ericsson and Simon is not appropriate for usability testing domain because the situations in usability testing is not very standardised as some factors depends on the test facilitator and the process of drawing an inference may be lost with context-free coding.

### 3.5 Data Analysis Approach: Thematic Analysis

The thematic analysis offers accessible and systematic procedures for generating themes and codes, which are minor units of analysis that capture interesting data features relevant to the research. Braun & Clarke (2017; 2006) offer an outline guide through the six phases of thematic analysis:

1. Familiarising yourself with your data: Transcribe data, read and re-reading the data, noting down initial ideas.

2. Generating initial codes: Coding interesting data features systematically across the entire data set, collating data relevant to each code.

3. Searching for themes: Collating codes into potential themes

4. Reviewing themes: Checking if the themes work in relation to the coded extracts (Level1) and the entire data set (Level 2), generating a thematic 'map' of the analysis.

5. Defining and naming themes: Generating clear definitions and names for each theme.

6. Producing the report: Relating the analysis to the research question and literature, producing a scholarly report.

### 3.5.1 Phase 1: Familiarising Yourself with Your Data

This phase serves as the foundation for the rest of the analysis. It entails being acquainted with the depth and breadth of the transcribed data through active repeated reading, searching for meanings, patterns, and so on (Braun & Clarke (2017).

In accordance with this phase, the author employs line-by-line analysis in the search for ideas and the discovery of patterns, languages used by participants, and their context of concern to aid in making sense of the data while keeping the underlying research question in mind.

In this phase the author also employs the use of note taking or recording ideas for coding that will be explored in subsequent phases.

### 3.5.2 Phase 2: Generating Initial Codes

In this phase the author starts to organise the data in a meaningful and systematic way. Since the author has gain some familiarity with the data and have generated an initial list of ideas about what is in the data and what is interesting about them. Overall, this phase involves the production of initial codes from the data.

The author was interested in participant experience with using think-aloud and the data was analysed with that in mind. So, it was an inductive analysis, so the author coded every piece of text using line-by-line coding to code every single line. The author make used of opening

coding; that is there were no pre-set codes, but codes were developed and modified during the coding process.

The author has an initial idea about the codes after phase one, which involves data familiarisation. For example, reviewing recording with a focus on task success rate and prompt for rich data; prompting participants for explanations and recommendations was an issue that kept coming up (in all the interview not just the extract) and they were truly relevant to the study's research question (iii) and (i) respectively.

As the author worked through the transcripts, new codes were generated, and previous codes were occasionally updated. This aids the author in forming some basic themes and codes. This was accomplished by working with pens and highlighters on hardcopies of the transcripts. Other tools can be useful as well, for example, Bree and Gallagher (2016) explained how to code and discover themes using Excel. However, qualitative data analytic software like NVivo can be extremely valuable, especially when dealing with enormous data sets.

Table 3 lists all the preliminary themes that have been identified, as well as the codes that go with them. Some of the utterance's categories were obtained from published usability studies, and others were derived or combined and redefined for them to be suitable for the current study. For instance, the utterance category: Reading, was defined according to Cook (2010) as "reading words, phrases, or sentences from the screen" and the utterance category: Action Description was defined according to Zhao and McDonald (2010) as "Describe what they were doing, trying to do or did". In this research both utterances: Reading and Action Description was combined to one single category called: Action Description and is defined as Read out text, sentences and links, describe what they were doing, trying to do or just did. In addition, utterance categories with * symbols are the same with McDonald (2013) and utterance categories with + symbols are the same with Zhao and McDonald (2010).

### Preliminary themes

| Emerging Themes | Sub-themes | Transcript |
|---|---|---|
| Think-aloud Usefulness | Expeditious<br>Task solving strategy<br>Insights, User experience<br>User preference | *"it's the thing that will give us insights into people's, um, expectations" [P1]* |
| Think-aloud as a Techniques | Constructive Interaction<br>CTA, RTA | *"I just ask them to just out loud what they, what they would normally say to themselves as they're doing the task." [P7]* |
| Instructions used during think-aloud | Explicit instructions<br>Neutral instructions<br>Reminders | *"I just let them know that there's no, um, there's no constraints on what they say." [P3]* |
| Implementation of practice sessions | No practice sessions<br>Occasional practice sessions<br>Practice sessions | *"Now I just, um, give them a pen and get them to take it apart of get them to use an unrelated product and just get them to think aloud." [P20]* |
| Tasks used during usability test | Representative tasks<br>Solvable tasks,<br>Task confusion<br>Task design, Task failure<br>Task scenario, Task skipped<br>Verbal tasks | *"So I would identify what the key, what the key functionality relates to either to the business and to what the business goals are for the product" [P14]* |
| Interacting with participants | Minimal interaction<br>No interaction | *"Um, so I'll try and keep my interactions pretty neutral and infrequent" [P11]* |
| Impact of test facilitator | Evaluator's led interaction<br>Task abandonment<br>Experience<br>Observation levels | *"Well, um, if you start asking people questions that might get them to, to change their strategy." [P6]* |
| Interventions during think-aloud session | Immediate prompt<br>Prompt for rich data<br>Stuck prompt<br>Reflective thinking | *"So if there was, if they made an injection or you know, that there was some element of surprise or confusion" [P5]* |
| Participants behavioural change | Exploration, Reactivity<br>Tasks, Culture | *"but I think it might change what they do" [P1]* |
| Participants explanations during think-aloud session | Method explanation<br>Testing process | *"And I make sure that I explain to them quite carefully what I want them to do." [P17]* |
| Evaluation based on project type | Commercial<br>Research, Formative<br>Summative, Project Budget<br>Remote | *"What I wouldn't be doing is transcribing it and doing a thematic coding." [P7]* |
| Client request during usability test | | *"or for what the client wants. And then I would make sure that the tasks I gave them related to that." [P12]* |
| Participant's characteristics during think-aloud session | Appreciation<br>Confirmation bias<br>Personality<br>Relaxation<br>Rep. participants<br>Rapport, Reassurance | *"You know, it might be that they're perfectionist and they want, they want to get everything right. And they want continue." [P21]* |
| The help participants need during think-aloud session | Clarification help<br>General help | *"Um, if there was something that happened that if they asked me a question, I would, I would answer it in such a way as not to steer them." [P8]* |
| Data Analysis activities | Tasks success rate<br>Session review<br>Metrics, Notetaking<br>Results discussions<br>Thematic coding<br>Utterance comparison | *"It will be, those tasks that I would focus my analysis activity on. And it would really be at the level of looking back at the videos and what people said and, or, or even checking the understanding that I've taken from the session." [P9]* |
| Think-aloud limitations | Participant dependent<br>Social desirability<br>Extends time, TA impact | *"so when I'm doing it, I tend to dry up, uh, that's the main limitations. Some people are not good at it. And I guess as well, it's a bit artificial" [P15]* |

*Table 3.1: Preliminary theme: REP – Representative*

### 3.5.3 Phase 3: Searching for Themes

The next phase involves sorting the different codes into potential themes and collating all the relevant data extracts within these themes. According to Braun & Clarke (2006) there are no fixed rules about what makes a theme. A theme is characterised by its significance. In this context there were few overlaps between the initial coding phase and this stage of identifying preliminary themes.

The author examined the codes and some of them clearly fitted together into a theme. For example, there were several codes that related to Interaction between the test facilitator and the participants, facilitated by the test facilitator and politely asking participants to move to the next tasks. The author collated these into an initial theme called the Impact of Test Facilitator.

At the end of this step the codes have been categorised into broader themes that corresponds to the study questions. Themes were mostly descriptive, describing patterns in the data that were pertinent to the study.

### 3.5.4 Phase 4: Reviewing Themes

In this phase the author review, modify and develop the preliminary themes that were identified in phase three. At this point It is useful to put together all the data that is relevant to each theme. Themes are reviewed and refined on two levels at this phase. The first stage entails evaluating the coded data extraction (reading all the collated extracts for each theme and consider whether they appear to form a coherent pattern). A similar procedure is used at level two, except this time it is applied to the full data set. The author considers the overall validity of each theme whether the chosen thematic map correctly reflects the meanings found across the whole data set.

The data associated with each theme was analysed to see if it supported the theme. The author also investigated if the themes might be applied to the entire dataset and how the themes work both in a single interview and between interviews.

For example, the author felt that the sub-theme: task under the preliminary theme: “Participants behavioural change”, did not really work well as a sub-theme under Participants behavioural change, as it overlaps with the theme: Tasks used during usability test. Hence, this was refined to Task Difficulty which capture an aspect of participants behavioural change which better captured what the participants were saying in the interview.

### 3.5.5 Phase 5: Defining & Naming Themes

This is the final refining of the themes, with the goal of determining the essence of each theme (Braun and Clarke, 2006, p.29). The author attempts to determine what each theme is about, as well as how the sub-themes interact with and connect to the primary theme, as well as how they relate to one another.

In this analysis: what are the practices and challenges of using the think-aloud protocol in the industry is an overarching theme that is rooted in the other themes.

### 3.5.6 Phase 6: Producing the Report

This phase involves the final analysis and write-up of the report. This involves telling the complicated story of the data in a way which convinces the reader of the merit and validity of the analysis. The author tells a clear, cohesive, logical, non-repetitive, and engaging explanation about the data's story within and between themes.

## 3.6 Utterance Categorisation

This study will utilise the think-aloud protocol, instructing participants to verbalize their thoughts as they interact with the system under test. The participants' utterances will be recorded, transcribed, and categorised based on the classification frameworks presented in

the literature. Following the categories established in the literature, we will categorize utterances into 'description', 'explanation', 'critical', and 'emotional' types for our usability study.

The researcher will use the categories of problem-related and non-problem-related utterances as defined by Boren & Ramey (2000), as well as the success-related, failure-related, and process-related categories from Hertzum and Holmegaard (2015) as a guide for the utterance categorisation in the studies presented in this thesis.

For instance, the utterance category “Reading” as defined according to Cook (2010), “Action Description” as defined by Zhao and McDonald (2010) and “Domain Knowledge” as defined by McDonald et al., (2013a). These combinations will provide a comprehensive understanding of the user experience, highlighting both interface issues and positive aspects. This will be emphasised within each study, as the categories of utterances may vary across different studies. See page 122 for details.

### 3.7 Reliability and Validity

Concurrent think-aloud began as a psychological technique for analysing expert chess success and designing cognitive models (Ericsson, 2006). However, its dependability and validity have long been questioned (Duncan, 1985; Smagorinsky, 2001; Ramey, et., 2006; Schooler, 2011; Fox et al., 2011).

Research evidence indicated that, asking people to think aloud while conducting tasks disrupts the participant's thinking process, resulting in a shift in task performance processes. This disruptive change is called reactivity (Schooler, 2008; Schooler, 2011; Fox et al., 2011). This has been discussed in detail in chapter 2 of this thesis.

Experiments should be replicable and produce similar results to be reliable. Human subjects, however, make it challenging to get repeatable results even with same subjects (Lazar et al.

2010, pp. 57). To be accurate, experiments must be repeatable and yield similar results. Human subjects, on the other hand, make it difficult to obtain recurring results even with the same subjects.

When a test participant completes a task several times, for example, the performance time varies, resulting in random errors. The resulting impact of these random errors can be reduced by increasing the sample size. However, larger sample sizes cannot reduce the bias that comes with systematic errors such as the errors that comes with test procedures, test facilitators behaviour, test environment, participants and equipment (Lazar et al. 2017). The use of an external test facilitator, as well as a team of facilitators from diverse backgrounds, will improve objectivity (Hughes 1999).

There are two ways introduced for the reliability of quantitative content analysis. Reliability can be established by the percentages of agreement among two or more coders, and the acceptable percentage is more than 80%. The other way of conducting reliability is to conduct Cohen's Kappa (Cohen, 1960; Ary, Jacobs, & Sorensen, 206; Riffle et al., 2005). Cohen's Kappa is used to assess the inter-rater reliability when coding categorical variables and refers to the proportion of observed units beyond that expected by chance alone. It is the measure of agreement between two individuals in coding quantitative data into categories. Numbers of agreements and disagreements between the raters can be entered into the statistical software to gain the value of reliability. A minimum value of 80% should be expected for an adequate reliability.

Validity refers to whether the usability test actually tests something that is relevant to the usability of real-world products outside of the lab. (Nielsen 1993, p. 169). According to Chi (1997), to establish the validity of the pattern coded from data, authors can code the data twice for an identified pattern, as the second coding process can be used to check validity with regards to subjective interpretation that may have occurred during the first process. For

instance, in this research, the authors can first list an initial pattern which preliminarily identified categories of cognitive engagement by reviewing the verbal data. The pattern with the preliminarily set categories can be validated by presenting no less than one verbal case as evidence to support the previously identified categories.

### 3.7.1 Coding Reliability for Study Three

The coding of the qualitative data was done by the author, within the framework of a PhD thesis, this was unavoidable. However, steps were taken to limit any possible bias. While the author did the coding for all 22 participants. The second coder was a lecturer from a different department who is well-versed in qualitative analysis and not been involved in the study other than the coding. The second coded data for six participants, both the author and the second coder go through the coded data to identify where there are agreement and disagreement. A new list of code where then produced that reflects the changes. The average kappa value of the agreement between the two coding was 0.84 (84%). An acceptable percentage is 80% Cohen, 1960; Ary, Jacobs, & Sorensen, 206; Riffle et al., 2005).

## 3.8 Ethics Based on Humans as Participants

A key component of research ethics is ensuring participants are treated with dignity and respect, and that their rights and wellbeing are safeguarded. Strict ethical rules that are intended to prevent participants from suffering physical, psychological, or emotional harm as well as to uphold their right to privacy and confidentiality must be followed when research involving human subjects is conducted (Dickson-Swift, James, & Liamputtong, 2008; Malacrida, 2007).

Informed consent is one of the fundamental tenets of research ethics, and it calls for participants to be fully informed about the purpose of the study, any possible risks or rewards, and their right to discontinue participation at any time (Resnik, 2021). In addition, without force

or undue influence, participants must freely and expressly accept to taking part in the research. Participant autonomy and decision-making capacity must be protected, and participants must be treated with respect and dignity, using informed consent (Dougherty, 2020). Hence, all three studies in this thesis gained full consent from all participants through an informed consent form, see details in appendix.

Confidentiality and privacy are two additional fundamental principles of research ethics. In order to avoid any potential harm or unfavourable outcomes that may result from the exposure of sensitive information, researchers must take all necessary precautions to secure the identity and personal information of participants. To gain the respect and trust of participants and to make sure they will continue to agree to take part in research in the future, confidentiality and privacy are crucial (Resnik, 2018). As detailed in the participants information sheets for all three studies in this thesis, information such as your gender, age, nationality, educational qualification, occupation and your use of the internet are kept anonymous and all Information were kept in a secure locked cabinet or a password protected computer within the University of Sunderland, see appendix B, participant information sheet on page 235 for details.

To ensure that the advantages outweigh the risks and that the research is conducted in a way that maximises the benefits while minimising the risks, research ethics also mandates that the possible risks and benefits of the study be properly assessed and weighed (Israel, 2015).

With respect to the studies in this thesis, it is the policy of the University of Sunderland that all research must be conducted in accordance with the University's Research Ethics Principles, Professional Codes of Practice and the law. All three studies in this thesis were subjected to appropriate ethical review by the Research Ethics Committee and gained approval using the online ethics review system. Hence, all three studies in this thesis were conducted in accordance with the University of Sunderland policy on research involving human participants and personal data.

## 3.9 Summary

The research goals and the need for a systematic approach to the analysis of the concurrent think-aloud and its procedural variables were presented in this chapter. Important experimental design issues were discussed, as well as qualitative research techniques for participants' utterances.

The next chapter investigates of the Impact of Task-types on the Reactivity of the Concurrent Think-Aloud in usability testing and found valuable results that will benefits both usability research and practice.

# CHAPTER FOUR

## 4.0 AN INVESTIGATION OF THE IMPACT OF TASK-TYPES ON THE REACTIVITY OF THE CONCURRENT THINK-ALOUD IN USABILITY TESTING

### 4.1 Overview

The first empirical investigation is presented in this chapter. The study investigates whether the act of thinking-aloud under classic administration procedures causes reactivity and if tasks performances with simple task-type such as fact tasks or difficult task-type such as sensemaking tasks have influence on the reactivity of the concurrent think-aloud by comparing the classic think-aloud with silent. The chapter finishes with recommendations for how the concurrent think-aloud could be improved after a discussion of the findings.

### 4.2 Motivation

Usability testing is an imperative aspect to the success of digital products and services. The Concurrent think-aloud (CTA) is one of the fundamental tools used for usability testing, in which users' verbalisation takes place simultaneously with their task performance. It is primarily used to understand users' task based cognitive processes and it is both time and cost effective (McDonald and Petrie, 2003). However, CTA is not without limitations, studies indicate that CTA procedures varies widely among practitioners Olmsted-Hawala, et., al (2010). Also, CTA has been known to cause reactivity: an artificial change, (enhanced or diminished performance) in task performance making the test no longer representative of real world use Fox, Ericsson, and Best (2011), this is problematic because it may alter the accuracy of task performance, thus leading to poor usability problem detection, low data reliability and validity, Ericsson and Simon, (1980, p. 27) stated that, "the accuracy of verbal reports depends on the procedures used to elicit them" and reactivity will occur when the established procedure is neglected. They also established guidelines to ensure the validity of CTA data elicitation, there are: (i) a neutral instruction with no specific information request (ii)

practice session and (iii) a neutral reminder to "keep talking" with no further enquiry. Studies show that practitioners don't follow these guidelines (McDonald, Zhao, and Edwards 2015). However, empirical demonstrations of reactivity within usability testing have shown mixed findings with some researchers finding evidence of reactivity and others finding none, despite modifying the CTA procedures to violate Ericsson and Simon's guidelines (Olmsted-Hawala, et., al 2010, Eger, et al., 2007). Thus, conclusions cannot be drawn to attest to whether reactivity occurs due to varying administration procedures within usability testing and therefore we must now consider its relationship to other test-based factors.

Usability testing is a task-based approach; thus, reactivity may not be solely attributed to CTA administration procedures as it may also depend upon other factor such as task type or level of difficulty. Indeed, when examining studies whose findings on reactivity do not concur, they often use different types of tasks. For example, tasks may involve fact-finding: to find out specific information, assessment tasks: to make judgement and opinion and sensemaking tasks which involves effort to understand relations in order to select a best course of action. Hence further investigation is required to examine task types, that may influence reactivity, thus threaten the validity and reliability of CTA.

Tentative evidence exists from a study which examined the impact of task difficulty on concurrent think-aloud (McDonald, McGarry, and Willis, 2013).

The results suggest that for difficult tasks CTA participants completed fewer tasks successfully whereas for easy tasks there were no differences in task performance among participants' when compared to a different variant of CTA. Thus, indicated that task difficulty may be a contributing factor to likely cause reactivity. As reactivity has impact on tasks either positively or negatively, hence, task type requires further investigations.

### 4.3 Research Aims

To bridge the gaps identified above on previous studies of CTA and reactivity, this study will investigate the impact of task type on the reactivity of the CTA within the context of usability testing. We investigate the following hypothesis:

The study uses two different task-types: fact and sensemaking tasks, we anticipated a difference in task performance with respect to task success due to levels of task difficulties and will also lead to an increase verbalisation during think-aloud.

#### 4.3.1 Hypothesis

The following assumptions were developed based on the literature review done in Chapter 2:

H1: Sensemaking tasks will affect tasks performance and decrease task success over fact tasks.

H2: Changes in performance will be more pronounced for sensemaking tasks and will lead to an increase in participant's verbalisation for CTA.

The study tasks will be completed in two different conditions: the classic think-aloud and the silent condition. Although, sensemaking task is more complex than fact task. We anticipated that this will not have an impact on the CTA protocol and thus will not influence reactivity of the CTA.

H3: The act of thinking-aloud under classic administration procedures and tasks performance with sensemaking task type does not cause reactivity within usability testing

### 4.3.2 Research Questions

This study will answer the following research questions:

**RQ1 (i)**: Does act of thinking-aloud under classic administration procedures causes reactivity?

**RQ1 (ii):** Does the use of different task-types: fact tasks and sensemaking tasks have impact on the reactivity of the concurrent think-aloud?

## 4.4 METHODOLOGY

This section detailed the design and test procedures on the present study. Permission to undertake the study was sought and approved from our University Research Ethics Committee.

### 4.4.1 Participants

Twenty volunteer participants were recruited from students at a university in the North-East of England, age range between 18 and 60 years. All participants were native English speakers in order to avoid difficulties with verbalisation during task performance. All participants reported to have a minimum of senior secondary educational qualification and a maximum of a master's degree qualification. All participants reported they were daily users of the internet and representative users of the test product as indicated by their responses to a user profile questionnaire.

### 4.4.2 Materials and Tasks

The Thomas Cook holiday travel website: (https://www.thomascook.com/) was selected because of its information-rich content and a state-of-the-art website with wide user base for representative users, also it's a type of product that participants would be happy to explore. During the formulation of the task two aspects were practically considered. Firstly, site

inspection and product walkthrough were conducted to determine area where participants might experience difficulties. Secondly, a considerable attention was given to the main features and functionality of the site, thus a clear goal for the tested product was established. The author developed ten tasks, five for fact and five for sensemaking tasks. Tasks were formulated in accordance with Dumas and Fox, (2009) which suggested three major areas where tasks should be structured around during the formulation of tasks for usability testing. These includes:(i) important tasks such as frequently performed tasks and basic tasks that involve the core functionalities of a product (ii) tasks that involve areas where the test evaluator preserved users might have difficulties and (iii) tasks that explore the product navigation and information architecture. Also, Redish and Dumas, (1999) suggest the use of scenario as it makes a task more realistic and because it eliminates the artificiality of the tasks by telling a reasonable and very short story. Hence, task was formulated with short story. For example:

#### 4.4.2.1 Fact task

You are planning your forthcoming wedding abroad and have decided to use Thomas Cook. Please find out the phone number of their weddings team.

#### 4.4.2.2 Sensemaking tasks

You wish to go on a Cruise to the Mediterranean in August 2017 for a duration of 10 to 14 nights and have chosen Southampton, England to be the port where the cruise ship will depart.

You DO NOT need a specific cruise line or cruise ship

i. You want to include airfare to and from Newcastle Intl. Airport, Newcastle
ii. You want a cruise that will visit at least 5 different ports
iii. You want a cabin with balcony and ocean view
iv. The total budget for the booking should be £4,450.

#### 4.4.2.3 Experiment Time Period

The usability study of the dynamic website: https://www.thomascook.com, was carefully designed to minimise the impact of potential site updates during the period of the study. The study was conducted over a period of two weeks. This duration was chosen as it allowed for enough data to be collected, while minimising the likelihood of significant changes being made to the website during the study.

To account for the dynamic nature of the website, the researcher diligently checked for updates before and after each usability session. This was crucial to ensure that the interface used by participants remained consistent throughout the study. As any potential updates that occurred would influence the usability of the site and hence, the findings of the study.

By taking these precautions, the researcher aimed to ensure that the results of the usability study accurately reflect the state of the website during the two-week experiment period, and there was no changes or updates on the site during the two weeks period. The researcher acknowledges that due to the dynamic nature of the site, some usability aspects may change post-study, however, the findings from the study provide a snapshot of the user experience during the controlled study window. The piloting approach for the first study is highlighted in section 3.3.8.

### 4.4.3 Questionnaires

This study included two additional questionnaires.

The TLX mental workload questionnaire designed by Hart and Stavenland (1988), was used to assess participants' mental workload. This is because the author wanted to know the difference between fact and sensemaking tasks with respect to task difficulty, performance, effort and frustration. This has been used in various think-aloud empirical studies such as Hertzum et al., (2009); McDonald et al., (2015).

The think-aloud After Scenario Questionnaire was a self-monitoring questionnaire designed by the author to find out about participants experience. A five-point Likert scale was used where participants indicate the extent of their agreement or disagreement.

### 4.4.4 Study Design & Sampling Method

A mixed design was adopted, where a within-subject design was used for task type (fact tasks and sensemaking tasks) and a between subject design was used on think-aloud. The rationale behind the adoption of a mixed design were, with a with-subject design introduces difficulties in ensuring that the group are equivalent, thus reducing the study statistical power: the likelihood that the study will detect reactivity if there is reactivity to be detected. Also, a between-subject design, risked carryover effects or practice effects caused by participant growing general familiarity with the study task set and test product and, prolong test time beyond what is reasonable. A mixed design eliminates the difficulties associated with using either method individually. In addition, a mixed design help to avoid carry-over effect, it is also efficient in terms of time and statistics.

The study consists of two groups, the first group “A” experimented with the classic think-aloud condition and group “B” was the silent condition, both groups performed fact-finding and sensemaking tasks. Participants were allocated randomly into one of the two groups, 10 participants for each group, see details in the diagram below.

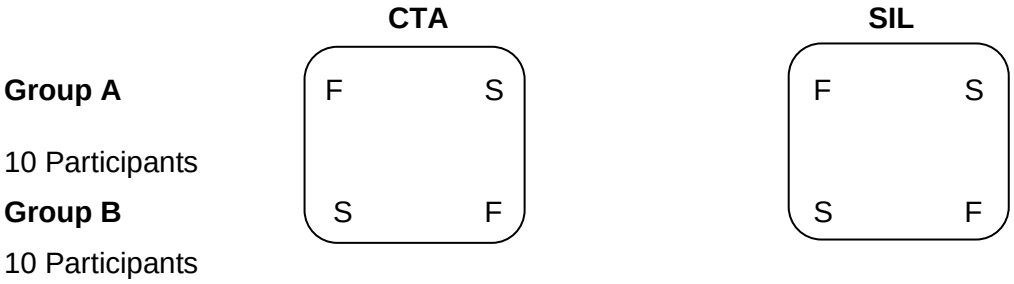



*Figure 2: Mixed design (where “CTA” denotes Concurrent think-aloud, “SIL” denotes Silent Working “F” denotes fact tasks and “S” denotes: sensemaking tasks)*

### 4.4.5 Study procedures

The test sessions took place in our usability laboratory and was facilitated by the author. Each session was conducted on a one-to-one basis and lasted around 50 minutes including introductions, debriefing and filling of all necessary consent forms. At the start test facilitator informed each participant on the that they would be completing 10 tasks with the website, 10 participants received the classic think-aloud instructions, which strictly followed Ericsson and Simon's guidelines, they practised thinking-aloud while disassembling a ball-point pen and the only interaction between the test facilitator and participant was to issue a neutral think-aloud reminder "please keep talking" if the participants fell silent for 15-20 seconds. The other 10 performed their tasks in silent. Participants were told that there were no time limits on task completion. They were asked to attempt all tasks and to only abandon a task if they felt that they had reached the point with a task where they would normally give up in real life. Participants were asked to read and ensure they understood the task requirements before they commerce with task performance. Then, a task booklet was handed to the participant and the test started. For both conditions, at the end of each task participants were asked complete the TLX workload scale to record their task experience. The test facilitator remained in the usability laboratory with participants and was seated a little way behind participant and to their right-hand side. At the end of the test while CTA participants were asked to complete two after test questionnaire, one to record their test experience and a second questionnaire to record their think-aloud experience. For the silent condition only one after test questionnaire was given to record their test experience. The test sessions were recorded (video and audio) including the screen using TechSmith Morae on the Tobi Studio Eye tracker machine.

## 4.5 Dependent Measures: Verbal Data

Verbal data: the type of utterances produced, and the number of utterances made. This study followed the verbal data analysis stages as set out by Chi, (1997), which was adopted by Zhao (2012, p36). Although not all Chi's recommended stages applies to this study, hence the

following stages that were applicable to this study is as follow: (1) To Develop or choosing a coding scheme; (2) To operationalise evidence in the coded protocols that constitutes a mapping to some chosen formalism; (3) Seeking pattern(s) in the mapped formalism.

The individual test sessions were transcribed and segmented into individual utterances for fact tasks and sensemaking tasks. The verbal utterances differ in length however, each represented a single unit of meaning. The individual utterances were annotated with the participant number and task number.

Some of the utterance's categories were obtained from published usability studies, and others were derived or combined and redefined for them to be suitable for the current study. For instance, the utterance category: Reading, was defined according to Cook (2010) as "reading words, phrases, or sentences from the screen" and the utterance category: Action Description was defined according to Zhao and McDonald (2010) as "Describe what they were doing, trying to do or did". In this research both utterances: Reading and Action Description was combined to one single category called: Action Description and is defined as Read out text, sentences and links, describe what they were doing, trying to do or just did. In addition, utterance categories with * symbols are the same with McDonald (2013) and utterance categories with + symbols are the same with Zhao and McDonald (2010).

### 4.5.1 Verbal Data Coding Reliability

The author coded the data twice using NVivo 11 with at least four weeks between the first and second coding for an identified pattern, as the second coding process is used to check validity with regards to subjective interpretation that may have occurred during the first process. For instance, in this research, the authors coded the data by attaching each utterance to an interpretative code, secondly the author goes through the coded data to Identify where there are agreement and disagreement and produce a list of the number of codes that changes with

reference to the previously identified categories; thirdly, a reliability score was done using the average Kappa value. The process of re-coding to assess the reliability corresponds with studies done in usability testing where utterances were coded twice (Hertzum, 2009; McDonald, Zhao and Edwards, 2013).

*Table 4:Coding scheme for utterance data*

| Utterance category | Definition | Example |
|---|---|---|
| Action Description | Read out text, sentences and links, describe what they were doing, trying to do or just did. | So am just reading, am looking for visa information, travel support customer advice, visa information on each [participant 9, task 3] |
| Action Explanation+ | Statement made to clarify the reason for before or after-action execution | Am going to look at customer services this could be a way to narrow it down [participant 4, task 2] |
| Expectation+ | Express expectations about what is/was going to happen or to find in a particular location, including indication of things that are counter to expectations | This phone number I presume it should be down at the bottom [participant 8, task 1] |
| Causal Explanation+ | Expressing or indicating a cause of difficulties | Okay I still can't find a flight why can't I do that? [Participant 10, task 6] |
| Result Evaluation+ | Giving justification to completed task(s) or describe what is on the website | Sri-Lanka is "AA", so it said a six-month visa is required [Participant 8, task 3] |
| User Experience+ | Expression of positive and negative feelings and experience caused by the site | So, I will just go back, my page just expired wonderful. [Participant 5, task 4] |
| Problem Indication + | Utterances indicating uncertainty, negative feeling or disapproval caused by website | Let's see all-inclusive I don't know what that means [participant 2, task 8] |
| Recommendation+ | Recommendation on how to improve the website or solution to difficulties experienced | So hmmm the party size doesn't make sense why don't it said your basket, [participant 8, task 7] |
| Impact^ | Explain what led to task difficulties, which may result to restarting the task or giving up | So am looking at this because am trying to make sense of it because is so much information coming at you [participant 10, task 5] |
| Domain Knowledge+ | Giving account of past experience with the similar website or type of tasks | I have tried to get one of this deal before like two weeks ago with a different travelling site [Participant 3, task 8] |
| Task confusion^ | Indicate confusion or misunderstanding of tasks | And the question do not said if I am going to pay by debit card or credit card [Participant 8, task 5] |
| Recollection | Discovering text, links, tasks steps they previously came across. | I already saw something at the bottom about visas, passport and visa [Participant 8, task 4] |
| Help Requests | Question(s) relating to tasks requesting to use certain website or system-based functions. | So, I guess am using can I use the main function on top of the webpage? [Participant 10, task 6] |

## 4.6 RESULTS

This section presents the results of this study in the following order: task performance, participants utterances and participant test experience.

### 4.6.1 Tasks Performance: Time on task

Table 4.1 presents the data for the mean time on tasks spend by participants in each think-aloud condition and for each type of task. There was a significant main effect of task ($F(1,18) = 38.576$, $p<0.001$). Overall, participants spend more time when completing Sensemaking tasks (mean= 329.97) than when completing fact tasks (mean= 150.58). There were no other significant main effects or interaction.

*Table 4.1: Time on tasks*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 266.34 | 110.92 | 393.60 | 101.16 |
| Fact Tasks | 131.96 | 24.64 | 169.18 | 45.09 |

*Significant difference obtained $p<0.05$

### 4.6.2 Number of clicks

Table 4.2 presents the data for the mean number of clicks made by participants in each think aloud condition and for each type of task. There was a significant main effect of task ($F(1,18) =63.243$, $p<0.001$). Overall, participants made more mouse clicks when complete Sensemaking tasks (mean = 44.21) than when completing fact tasks (mean= 12.99). There also a significant main effect of think-aloud ($F(1,18) = 5.191$, $p<0.035$). Overall, participants made more mouse clicks when thinking aloud (mean= 33.73) than when working in silence (mean = 23.47). There was no interaction between task and think aloud.

*Table 4.2: Number of clicks*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 39.08 | 18.89 | 18.12 | 10.06 |
| Fact Tasks | 7.86 | 3.53 | 14.85 | 45.09 |

*Significant difference obtained p<0.05

### 4.6.3 Additional pages

Table 4.3 presents the data for the mean number of additional pages navigated by participants in each think-aloud condition and for each type of task. There was a significant main effect of task ($F(1,18) = 24.701$, $p<0.001$). Overall, participants navigated more pages and carry out more scrolling when completing Sensemaking tasks (mean= 4.85) than when completing fact tasks (mean= 2.11). There were no other significant main effects or interaction.

*Table 4.3: Number of additional pages*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 3.80 | 2.41 | 5.90 | 3.52 |
| Fact Tasks | 1.30 | 0.96 | 2.92 | 2.12 |

*Significant difference obtained p<0.05

### 4.6.4 Correct tasks

Table 4.4 presents the data for the mean number of correct tasks completed by participants in each think-aloud condition and for each type of task. There was a significant main effect of task ($F(1,18) = 41.705$, $p<0.001$). Overall, participants completed more fact tasks (mean= 0.67) than Sensemaking tasks (mean= 0.22). There were no other significant main effects or interaction.

Table 4.4: correct tasks

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 0.74 | 0.26 | 0.20 | 0.21 |
| Fact Tasks | 0.60 | 0.36 | 0.24 | 0.12 |

*Significant difference obtained p<0.05

### 4.6.5 Number of Abandon (Incomplete) Tasks

Table 4.5 presents the data for the mean number of abandon tasks completed by participants in each think-aloud condition and for each type of task. There was a significant main effect of task (F (1,18) = 10.440, p<0.005). Overall, participants abandon more Sensemaking tasks (mean= 0.41) than fact tasks (mean= 0.13). There were no other significant main effects or interaction.

*Table 4.5: Number of abandon tasks*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 0.42 | 0.23 | 0.41 | 0.20 |
| Fact Tasks | 0.16 | 0.15 | 0.01 | 0.31 |

*Significant difference obtained p<0.05

### 4.6.6 Partly Completed Tasks

Table 4.6 presents the data for the mean number of partly completed tasks completed by participants in each think-aloud condition and for each type of task. There was a significant main effect of task (F (1,18) = 51.761, p<0.001). Overall, participants completed more fact tasks (mean= 0.01) than Sensemaking tasks (mean= 0.36) which were partly completed and, in some cases, abandoned. There were no other significant main effects or interaction.

*Table 4.6: Partly completed tasks*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 0.34 | 0.18 | 0 | 0 |
| Fact Tasks | 0.02 | 0.06 | 0.38 | 0.22 |

*Significant difference obtained p<0.05

### 4.6.7 Number of Incorrect Solutions on Tasks

Table 4.7 presents the data for the mean number of incorrect solutions on tasks completed by participants in each think-aloud condition and for each type of task. There was a significant main effect of task (F (1,18) = 51.761, p<0.018). Overall, participants produce more incorrect solution on fact tasks (mean= 0.20) than Sensemaking tasks (mean= 0.03) which are mostly abandon. There were no other significant main effects or interaction.

*Table 4.7: Number of incorrect solutions on tasks*

| | Silent Working | | Classic Think Aloud | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Sensemaking | 0.06 | 0.18 | 0 | 0 |
| Fact Tasks | 0.22 | 0.23 | 0.18 | 0.22 |

*Significant difference obtained p<0.05

### 4.6.8 Mental Workload

Table 4.8 summarises the TLX results for both conditions: classic think-aloud and silent condition and task types: sensemaking and fact tasks respectively. Significant results are marked with asterisk. There were significant differences for the following subscales (independent t-test): Mental Demand (t= .315, df = 18, p=.041); Performance (t= 1.529, df = 18, p=.046); Effort (t=.641, df = 18, p=<0.014).

*Table 4.8: Mental workload*

| | Silent Working | | | | Classic Think Aloud | | | |
|---|---|---|---|---|---|---|---|---|
| | Sensemaking | | Fact tasks | | Sensemaking | | Fact tasks | |
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| Mental demand* | 59.30 | 21.61 | 44.40 | 20.29 | 62.30 | 13.68 | 46.80 | 12.99 |
| Temporal demand | 40.10 | 27.53 | 37.70 | 26.47 | 41.00 | 18.44 | 35.80 | 16.14 |
| Performance* | 44.80 | 14.19 | 67.30 | 23.84 | 48.70 | 19.48 | 80.30 | 12.40 |
| Effort* | 59.90 | 24.92 | 46.90 | 22.26 | 70.60 | 16.57 | 52.20 | 13.68 |
| Frustration | 44.12 | 26.54 | 28.20 | 20.29 | 63.60 | 21.98 | 19.88 | 13.44 |

Scale: 0 (very low) to 100 (very high) *: significant difference obtained $p<0.05$

### 4.6.9 Verbal Utterance

Table 4.9 presents the data for verbal utterances in the concurrent think-aloud phase, since the study compares CTA with Silent condition only the verbal utterances for the think-aloud condition was presented. Overall, the verbal data indicated that "action description", "action explanation" and "causal explanation" yielded more verbal responses. These utterances indicate participants verbalising their thoughts associated with difficulties' when using the website. Also, problem indication was the fourth highest categories, indicating uncertainty caused by the website.

*Table 13: Verbal utterance*

| **Utterance Categories** | **CTA (Sum)** |
|---|---|
| Action Description | 367 |
| Action Explanation | 264 |
| Causal Explanation | 95 |
| Domain Knowledge | 1 |
| Expectation | 12 |
| Help Request | 10 |
| Impact | 26 |
| Problem Indication | 96 |
| Recollection | 7 |
| Recommendation | 6 |
| Result Evaluation | 20 |
| Tasks Confusion | 18 |
| User Experience | 36 |
| Total | 956 |

Table 4.9: Number of Utterances produced in the concurrent think-aloud phase

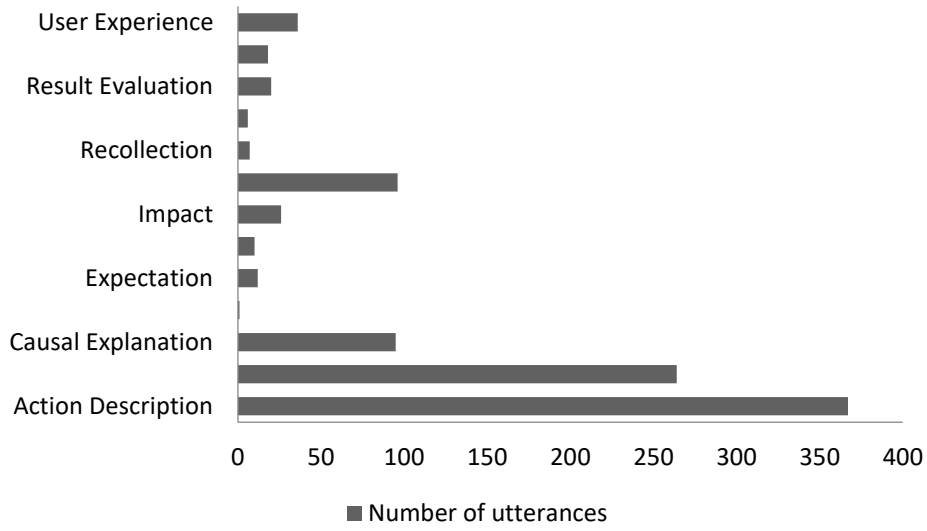



*Figure 3: Utterance categories produced in the concurrent think aloud phase*

## 4.7 DISCUSSION

The primary aim of this study was to explore the Impact of task-types on the reactivity of CTA and the impact of think-aloud on task-type: sensemaking and fact tasks. Findings from this study support results obtained from previous studies by examining the impact of tasks-type on the concurrent think-aloud. More importantly, it extends pervious work through the identification of the impact of think-aloud on task-types and the reactivity of CTA as we investigated the effect of sensemaking tasks on performance, time on task, number of clicks, correct tasks, number of abandon tasks, partly completed tasks and no of incorrect solution on tasks and task success. We will consider the limitation of our study and final conclusions and claims will be reached based on the present study and the corresponding results.

### 4.7.1 Task Performance

In terms of performance data, findings from our study indicated that, participants spend more time when completing sensemaking tasks than when completing fact tasks and was also long during think-aloud than when participant perform task in silence. There was a significant main effect of task ($F(1,18) = 38.576$, $p<0.001$). Overall, participants spend more time when

completing Sensemaking tasks (mean= 329.97) than when completing fact tasks (mean= 150.58). There were no other significant main effects or interaction. See details in table 4.1: Time on tasks, page 98. The additional time for sensemaking tasks and during think-aloud session may be attributed to tasks complexity and the fact that verbalisation is a slower process compared to think. This finding is in line with (Hertzum, Hansen and Andersen, 2009) who find out that task completion times were longer during thinking aloud than when participants performed tasks in silence. Although, less time is spent on fact tasks compare to sensemaking tasks, this may be due to the simplicity of fact tasks compared to sensemaking tasks with multiple route that led to task completion and lower a priori determinability: a measure of the extent to which a participant can deduce the required task, find necessary information and recognise the required information based on the task requirements. Evidence from study has shown that time spent on tasks is not because participants had to think-aloud during task performance but could be linked to the level of task difficulty.

In terms of the numbers of mouse clicks, participants made more mouse clicks when complete difficult tasks such as sensemaking tasks than when completing less difficult tasks such as fact tasks. Participants also made more mouse clicks when thinking aloud than when working in silence. This finding support evidence from previous study Van de haak (2003) which indicated that the task to concurrently think aloud caused more extra (observed) problems than it revealed in participants verbalisations, although, their study was not related to task-types in terms of tasks difficulties.

In terms of navigation, participants navigated more pages and carry out more scrolling when completing Sensemaking tasks than when completing fact tasks. This finding is similar with study conducted by (McDonald, McGarry and Willis, 2013) who indicates that for low difficulty tasks, there was no difference in task success or in the number of link traversals. However, for difficult tasks participants completed fewer tasks successfully and engaged in more link traversals, although their study investigated difference in navigation performance between explicit explanation-based think-aloud instruction and the classic think-aloud protocol.

In terms of correct tasks, participants completed more fact tasks than sensemaking tasks and abandon more sensemaking tasks than Fact tasks. Overall, participants completed more fact tasks than sensemaking task. Also, in terms of incorrect solution, participants produce more incorrect solution on fact tasks than sensemaking tasks, although, this may be as a result of tasks abandonment during the performance of sensemaking tasks.

Additionally, in terms of performance data, our result suggests that there weredifference between sensemaking tasks and fact tasks with respect to the data obtained from each task type as changes in performance were more pronounced during the sensemaking tasks which also led to an increase in participants' verbalisation during the concurrent think-aloud session. Therefore, we accept H2: changes in performance will be more pronounced for sensemaking tasks and will lead to an increase in participant's verbalisation for CTA. This finding is in line with (McDonald, McGarry and Willis, 2013) who also found that for more difficult tasks there were differences as participants in the classic condition completed fewer tasks and engaged in more link traversals.

### 4.7.2 Reactivity

This study demonstrated the 'non-reactivity' of the concurrent think-aloud, the result accords with (Ericsson and Simon, 1980; Van den Haak, de Jong and Schellens, 2009; Ericsson and Fox, 2011; McDonald and Petrie, 2013; McDonald, Zhao and Edwards, 2015) and discords with (Wright and Converse, 1992; Van Den Haak, De Jong and Jan Schellens, 2003; Eger *et al.*, 2007; Olmsted-Hawala *et al.*, 2010) who's studies found the CTA to be associated with reactivity. Therefore, we accepted H3: the act of thinking-aloud under classic administration procedures and tasks performance with sensemaking task type does not cause reactivity within usability testing.

Although, participants spend more time when completing sensemaking tasks than when completing fact tasks, there were no other significant interactions. Also, participants made more mouse clicks when thinking-aloud than when working in silence this maybe as a result

of the tasks complexity as evidence from the result further indicated that participants navigated more pages and carry out more scrolling when completing sensemaking tasks than when completing fact tasks. This finding concurs with Ericsson and Simon, (1993) which stated that "the accuracy of verbal reports depends on the procedures used to elicit them" and reactivity will occur when the established procedure is neglected and (Fox, Ericsson and Best, 2011) indicates that, a deviation from established guidelines often induce reactivity.

When looking at previous study that has been conducted on CTA, our findings concurs with (McDonald, McGarry and Willis, 2013) which suggested that for difficult tasks CTA participants completed fewer tasks successfully whereas for easy tasks there were no differences in task performance among participants' when compared to a different variant of CTA. The most plausible explanation of our results and previous studies concerning the concurrent think-aloud and reactivity may be partly cause by methodological irregularities as suggested by Hertzum, Hansen and Andersen, (2009).

### 4.7.3 Verb utterances

Participants did not rate the experience of their think-aloud negatively, indicating that participant did not perform worse on sensemaking tasks because they had to think-aloud and carry out tasks simultaneously and think-aloud did not affect their speed of working and their task focus. Although, this is contrary to findings from Van Den Haak (2003) which suggested that the cognitive load of the tasks combined with the extra task to think-aloud appears to have had a negative effect on both the participant's verbalisations and their task performance. However, (Hertzum, Hansen and Andersen, 2009) suggested that the act of thinking aloud alone is unlikely to cause reactivity; except for methodological irregularities.

## 4.8 Limitation and future work

There are several limitations of this study. First, the author functioned as both the test facilitator and the data coder. Ideally, different individuals would have performed the coding activities. To mitigate the potential bias this introduced the following measures were taken: (i) there was a delay of three weeks between data collection and transcription and a further five weeks for subsequent qualitative analysis; (ii) the second author crossed checked all qualitative data without knowledge of which data belongs to a particular TA placement.

Secondly, this study makes use of a small number of participants, although we did have a small sample size, the qualitative data that was gathered from the study add a valuable insight to our knowledge of the concurrent think-aloud and reactivity within usability testing.

Future research should investigate the value of level 3 verbalisation

## 4.9 Summary

Ericsson and Simon, (1980, 1993) stated that, “the accuracy of verbal reports depends on the procedures used to elicit them” and reactivity will occur when the established procedure is neglected. Some argue just the act of thinking aloud during tasks performance will cause reactivity (Van den Haak, de Jong and Schellens, 2004). Others argue that it depends on the elicitation procedures used (Fox, Ericsson and Best, 2011). Likewise, Hertzum, Hansen and Andersen, (2009) highlighted that the act of thinking aloud alone is unlikely to cause reactivity, except for methodological irregularities.

Evidence from this study demonstrates that the act of thinking aloud under classic administration procedures does not cause reactivity within usability testing and task type does not have impact on the reactivity of the concurrent think-aloud, although, sensemaking task lead to an increase in mental demand and effort.

This provides researchers and practitioners with a better understanding of the conditions that may affect the concurrent think-aloud protocol in terms of reactivity, the reliability of test data and provide valuable recommendations to help usability practitioners guide test design. This study has no evidence which indicate a change in participant's behaviour thus, indicating non-reactivity of the CTA.

# CHAPTER FIVE

## 5.0 THE IMPACT OF TASK-TYPE ON TWO DIFFERENT THINK-ALOUD PROTOCOLS IN USABILITY TESTING

### 5.1 Overview

This chapter presents the second empirical study of this PhD research. It explores the impact of task-type on using two different think-aloud protocols, the classic and the explicit instruction. The former follows Ericsson and Simon's recommendations regarding the use of think-aloud. In contrast, the latter requires explanations from users about their thoughts and navigation process.

The study focuses on issues relating to the working habits of usability practitioners with a focus on a significant aspect of divergent practice, such as test facilitators' use of instruction during usability testing. Findings obtained from the study was discussed, the chapter concludes and presents possible recommendation and future research.

### 5.2 Motivation

Studies within usability testing have documented divergent practice in the use of think-aloud instructions (Boren & Ramey 2000). The procedures and practices used indicated discrepancies between Ericsson and Simon's established model and reported practices during usability tests (Boren & Ramey 2000; McDonald et al. 2012).

Empirical demonstration of the use of instructions within usability testing has shown mixed findings. Some findings indicate that explicit instruction improves users problem solving strategies (Gerjets et al., 2011) and improves task performance (Wright & Converse, 1992). On the contrary, a study conducted by McDonald and Petrie (2013) shows that explicit instruction led to an increase in within and between page navigation and scrolling activity. Others reported that explicit instruction did not improve performance. However, increase

participants' mental workload and make participants more critical about the tested product (Zhao et al., 2014).

Indeed, the core idea behind usability testing is to observe real users try to accomplish actual tasks to collect accurate data; therefore, the task is a crucial part of usability testing. A study conducted by McDonald et al., (2013a) investigated whether an explicit explanation-based think-aloud instruction leads to differences in navigation performance over the classic think-aloud method. Findings indicated that participants on the classic condition completed fewer tasks successfully. Whereas for easy tasks, there were no differences in task performance among participants when compared to the explicit condition.

Consequently, the disparity in performance with the use of instructions within usability testing is unclear. Therefore, there is a need to consider other test-based factors, or could it be task-types?

## 5.3 Research Aims

This study builds upon previous studies within the field of usability testing to investigate the impact of task-type on two different think-aloud protocols and its effect on participant task performance, test experience and verbalisation by comparing the classic think-aloud, explicit instruction and silent working with fact and assessment tasks.

### 5.3.1 Hypothesis

The results and analyses that will be obtained from this study are intended to test the following research hypotheses. Based on Ericsson and Simon, (1993) protocol analysis which suggest that an explicit instruction will lead to an improved task performance. Hence, this study anticipated that:

H1: Participants on the explicit condition are more likely to perform better in terms of task success, mouse clicks and number of additional pages on both fact and assessment task.

Previous Study (McDonald, and Edwards 2014) suggest that explicit instruction can lead to higher mental workload, thus:

H2: The use of explicit instruction with fact and assessment task might lead to an increase in participant mental workload.

Studies conducted by (Boren and Ramey, 2000; Hertzum et al., 2009; Nørgaard and Hornbaek, 2006) suggest that test facilitators' intervention when using the think-aloud techniques during usability test sessions by instructing participants to comment on specific instances in order to obtain desired results, hence, this study anticipated that:

H3: The use of explicit instruction with fact and assessment may lead to an increase in participants verbalisation compared to the classic condition

H3b: Utterances will be more pronounced for utterance categories such as user experience and expectations compared to the classic condition.

### 5.3.2 Research Question

This study will answer the following research questions:

**RQ2(i)**: What is the impact of task performance on the use of fact and assessment task with the classic think-aloud, explicit instruction or silent within usability testing?

**RQ2(ii)**: Does explicit instruction lead to high mental workload over classic think-aloud and Silent?

**RQ2(iii)**: Does explicit instruction lead to an increase in relevant explanatory utterances in terms of user experience and expectations?

## 5.4 METHODOLOGY

This section detailed the design and test procedures on the present study these includes participant, study design, material and task.

### 5.4.1 Participants

Sixty volunteer participants were recruited from student at the University of Sunderland, 45 males and 15 females, aged range between 18 and 35 years, with a mean of 29 years. Internet experience ranged from 6 years to 18 years, with mean value of 10 years. All participants were fluent in English to avoid difficulties with verbalisation during think-aloud. Participants reported to have a minimum of senior secondary educational qualification and daily users of the internet. A user profile questionnaire was used to check that participants are representative users for the test product. See participant detail in appendix B6.

### 5.4.2 Material and Tasks

The nexus public transport services website: https://www.nexus.org.uk/ was selected because of rich information base which encompasses transport services on train, buses, ferry that gets people to work, or takes children to school, treating people to a day out shopping or a family trip to the coast or a museum.

#### 5.4.2.1 Task Derivation and Piloting

Usability testing is a task base approach and one effective way to identify usability problems is by observing users as they carry out series of tasks. All tasks were formulated in accordance with the suggestion of (Dumas and Fox, 2009, p.233). See section 2.6.1 for details on task derivation.

#### 5.4.2.2 Task Definition

| | |
|---|---|
| **Fact tasks** | These are type of tasks in which participants gathered information that is explicitly available on the web sites.<br><br>Task example: Find out if you can park your bicycle at the Monument Metro station and the numbers of bicycle racks |
| **Assessment tasks** | These are type of tasks in which participants gathered information and based on this information formed an opinion.<br><br>Task example: You would like to use the live travel map to see the next available Go North-East bus 700 from Sunderland University Travel Hub and download the timetable. How would you accomplish this task? |

*Table 5: Definition of fact and assessment tasks according to Spool et al. (1999), with task sample from the present Study*

### 5.4.3 Questionnaires

This study included two additional questionnaires, the TLX mental workload questionnaire designed by Hart and Stavenland (1988), was used to assess participants' mental workload. This is because the author wanted to know the difference between fact and assessment tasks with respect to task difficulty, performance, effort and frustration. This has been used in various think-aloud empirical studies such as Hertzum et al., (2009) and Mcdonld et al., (2015). The think-aloud After Scenario Questionnaire was a self-monitoring questionnaire designed by the author to find out about participants experience. A five-point Likert scale was used where participants indicate the extent of their agreement or disagreement.

### 5.4.4 Study Design

This study makes use of a mixed design: where a between subject design was used for the think-aloud conditions, participants were assigned to either classic think-aloud, explicit instruction or silent condition. A within-subject design was used for task-types: fact tasks and assessment tasks, all participants carried out task performance with both fact and assessment tasks.

The rationale for using a mixed design is to eliminate the difficulties associated with using either a within-subject or between-subject method individually and it is time efficient. In relation to this study, a within-subject design, risked carryover effects or practice effect caused by participant growing general familiarity with the study tasks-set and test product and it will also prolong test time beyond what is reasonable. Although, a between subject design is associated with individual difference factor, this was mitigated by a random placement of task types and increased number of participants which was twenty for each condition. The rationale for choosing the adopted explicit think-aloud Instruction for this study is based on analysis of the think-aloud literature and previous think-aloud studies within usability testing, see table 5.2 for think-aloud instructions.

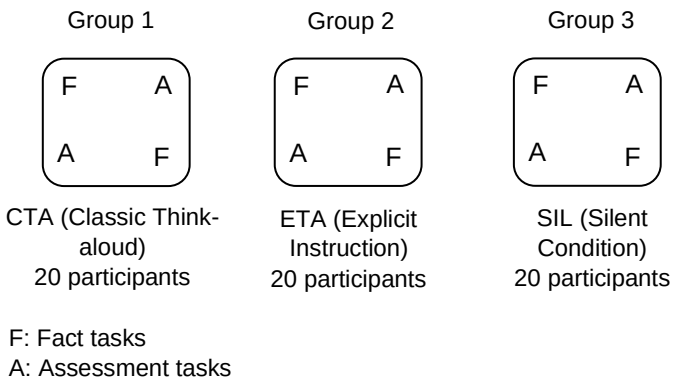



Figure 4: Diagrammatic representation of study design

The study consists of three groups, the first “group-1” experimented with the classic think-aloud condition, “group-2” experimented with explicit instruction and “group-3” was the silent condition. All groups performed fact-finding and assessment tasks; participants were allocated randomly into one of the three groups of 20 participants.

### 5.4.5 Independent variables: CTA, Explicit instruction & Silent condition

#### 5.4.5.1 Dependent variables:

i. Time on task: The amount of time a participant is actively engaged in performing a specific task. The estimated time on task will start when participants start reading the task and their first sight on The Nexus website homepage to when participants announced they have completed the task and write it on the answer booklet.

ii. Mouse Click: number of times the participant clicked inside an AOI or AOI group.

*In the context of a usability study, AOI stands for "Area of Interest", which is a specific part of a user interface that the researcher wants to focus on during the study. The initial AOIs for this study include buttons, text boxes, images, menu bars and the "Add to Cart" button that users might interact with or look at the Nexus website.*

*The researchers intended to track various metrics within these AOIs, such as the amount of time users spend looking at them, the number of clicks they receive, or the sequence in which users look at different AOIs. This information can provide valuable insights into how users interact with the Nexus website, which elements attract their attention, and which elements may be causing confusion or difficulty.*

*This is especially common in eye-tracking studies. However, due to the eye tracker calibration issues with some of the participants this metrics was neglected.*

iii. Successfully completed tasks: This will be determined by comparing participants’ answers on the answer booklet with the correct answers

iv. Number of abandon tasks: incomplete task, where participants decide to leave a task

v. Number of incorrect solutions: Where participants provide incorrect solution when compared with the correct answer

vi. Number of additional pages: This will be determined by between page mouse clicks

vii. Partly completed tasks: Tasks that are partly completed, this will be determined by facilitator’s observation of participant’s behaviour

viii. Verbal data: the types of utterances produced by participants during explicit instruction and CTA condition.

ix. Perceived mental workload: Mental workload will be measured using the NASA TLX workload questionnaire.

x. Participants' test experience: After test questionnaire will be use to collect data with regards to participants test experience.

### 5.4.6 Study procedures

Upon arrival, participants were greeted, and the test facilitator explained the purpose of the study which was to evaluation of Nexus website. All sessions were conducted on a one-to-one basis and permission to run the study was obtained from the university's ethics committee. Participants were asked to complete a consent form; and a user profile questionnaire for all three conditions: classic think-aloud, Explicit Instruction (explicit instruction) and silent. Test facilitator told participants not to turn to him for assistance and to pretend as if the facilitator is not there. Participant received an oral instruction read by the test facilitator from a paper to ensure consistency in conducting the test sessions. See table 5.2 for details of the instructions used for this study.

Table 5.2 Test instructions for classic, explicit instruction and silent condition

| classic think-aloud condition | Explicit Instruction | Silent |
|---|---|---|
| Participants on the classic think-aloud condition were given oral instructions to think aloud while performing tasks and make use of the instruction set out by Ericsson and Simon when conducting usability test with the classic think-aloud protocol such as “I want you to say out loud everything that you say to yourself in silent” if you fall silent for 15-20 seconds, I will remind you to “keep talking”. | Participants on the explicit instruction conduction make use of the following explicit instruction during think-aloud to request for explanations from users: "I would like you to think-aloud. I would like you to tell me the things that you like the things that you dislike or finding confusing about the site”. To ensure consistency and clarity, the test facilitator asked participants on the explicit condition to read out loud the instruction which was only written on the first task booklet before commencing with task performance. | Silent condition, participants were asked to perform tasks without verbalising their thoughts, they were instructed to solve the task and report their answers to the test facilitator upon completion, and this is similar to how users make use of a product when they are not undergoing a usability evaluation test. |

*Table 5.2: Test instructions for classic, explicit instruction and silent condition*

Table 5.2 shows the different test instructions used in this study, the instruction used for the Silent condition indeed acts as the baseline or control condition. As indicated in the table above, participants in this condition were asked to interact with the Nexus website and perform tasks without verbalising their thoughts, mirroring the usual way most users interact with a website in a non-testing situation.

The other two conditions, Classic and Explicit, involve variations of the think-aloud protocol where participants are instructed to verbalise their thoughts, as indicated in the table above. However, to understand the impact of these think-aloud protocols, its necessary to compare the results from these conditions against a baseline. This is where the Silent condition comes in. By comparing the results of the “Classic Think-aloud” and “Explicit instruction” conditions against the “Silent” condition, the study can assess how much the act of

verbalising using different think-aloud variant can affects the user's interaction with the Nexus website, their task performance and overall user experience.

For classic think-aloud and the explicit instruction conditions the only difference was the instruction, other test procedures were kept constant. After completing questionnaire, participants were asked to practice thinking-aloud using a neutral task: disassembling a ball-point pen and put it back together, while thinking aloud. The test facilitator reminded participants that it is the Nexus website that is been tested and not them. For all three conditions, tasks were given to participants in printed booklets and answer booklets, and at the end of each task participants were asked complete the TLX workload questionnaire. During the think-aloud the facilitator only speaks to remind participants to "keep talking" if they fell silent for 15-20 seconds.

At the end of the test sessions, two questionnaires were given to the classic think-aloud and explicit instruction conditions to record their test experience and think-aloud experience. While for silent condition only one questionnaire was given to participant to record their test experience. The test sessions were recorded (video and audio) including the screen using TechSmith Morae on the Tobi Studio Eye tracker machine.

### 5.5.1 Verbal Data: coding reliability

The author coded the data twice using NVivo 11 similar with the first study in this thesis with four weeks between the first and second coding for an identified pattern, as the second coding process is used to check validity with regards to subjective interpretation that may have occurred during the first coding process.

The four weeks between the first and second coding which led to the segmentation of five utterances and the correction of twenty-five codes. The average kappa value of the agreement between the two coding was 0.86 (86%). An acceptable percentage is more than 80% (Cohen, 1960; Ary, Jacobs, & Sorensen, 206; Riffle et al., 2005). The process of re-coding to assess

the reliability corresponds with studies done in usability testing where utterances were coded twice (Hertzum, 2009; McDonald, Zhao and Edwards, 2013).

| Utterance category | Definition | Example |
|---|---|---|
| **Action Description** | Read out text, sentences and links, describe what they were doing, trying to do or just did. | *“Am scrolling, am looking, am clicking on the bus hmm tab on the top to see if I can find something” [ETA1]* |
| **Action Explanation+** | Statement made to clarify the reason for before or after-action execution | *"Okay am going to click on the top one because I guess that’s hmm the address has several different addresses options" [ETA2]* |
| **Expectation+** | Express expectations about what is/was going to happen or to find in a particular location, including indication of things that are counter to expectations | *"I would love it to be a little thing just like travel help bar or something like that". "I think there should be something to tell me what the zones look like on the metro" [ETA18]* |
| **Causal Explanation+** | Expressing or indicating a cause of difficulties | *"Here the information is a little bit easier to grasp". "And how much is the ticket going to be now" [CTA10]* |
| **Result Evaluation+** | Giving justification to completed task(s) or describe what is on the website | *"So am guessing from north shields to South Shields on a Sunday is 6 o clock" [ETA10]* |
| **User Experience+** | Expression of positive and negative feelings and experience caused by the site | *"Nothing stands out everything is like in this grey colour" [ETA3]* |
| **Problem Indication +** | Utterances indicating uncertainty, negative feeling or disapproval caused by website | *"So, it doesn’t show me the nearest metro but hmmm let’s take a look again" [ETA19]* |
| **Recommendation+** | Recommendation on how to improve the website or solution to difficulties experienced | *"Nope here latest news, information, and alert no I am looking for journey planner that should be on the home page". "Anyway, I feel like the pop phone number should have been on pop help" [ETA8]* |
| **Impact*** | Explain what led to task difficulties, which may result to restarting the task or giving up | *"So am going to ignore that and am going to go back again to the metro home page". "Start your journey no, planned works let’s see, no that’s not what am looking for, so I will go back" [ETA4]* |
| **Domain Knowledge+** | Giving account of past experience with the similar website or type of tasks | *"I have not use this for about 11 year" [ETA13]* |
| **Task confusion*** | Indicate confusion or misunderstanding of tasks | *"Has she got pop card, has she not? So, we assume that she hasn’t got a pop card because it didn’t say she has". "You didn’t state the date, so I will assume is Monday to Friday" [ETA17]* |
| **Recollection** | Discovering text, links, tasks steps they previously came across. | *"In this section how to guides, pay zone I have seen this before". "Okay when last did I see pop card from home". "I actually search that before" [CTA14]* |
| **Help Requests** | Question(s) relating to tasks requesting to use certain website or system-based functions. | *"Can I get the postcode from Google?". "Did you say you are not allowed to use search?" [ETA17]* |

*Table 5.3 Utterance categories and their definitions for classic and explicit conditions*

Some of the utterance categories were obtained from published usability studies, and others were derived or combined and redefined for them to be suitable for the current study. For instance, the utterance category: Reading, was defined according to Cook (2010) as "reading words, phrases, or sentences from the screen" and the utterance category: Action Description was defined according to Zhao and McDonald (2010) as "Describe what they were doing, trying to do or did". In this research both utterances: Reading and Action Description was combined to one single category called: Action Description and is defined as Read out text, sentences and links, describe what they were doing, trying to do or just did. In addition, utterance categories with * symbols are the same with McDonald et al., (2013a) and utterance categories with + symbols are the same with Zhao and McDonald (2010).

The rationale behind this was due to the type of utterances obtained from the current study where most participants' reading out text then try to describe their activities and carry out task execution process by reading out links in relation to the underlying task at hand. See table 3 above for details of utterance category, definition and example from the study.

## 5.6 RESULTS ANALYSIS

### 5.6.1 Tasks Performance

An independent t-test found no impact of task on think-aloud as shown in table 5.4.

Although, there was a significant main effect of task ($F(1,57) = 1288.372$, $p<0.001$), participants spend more time during ETA than on CTA when completing Assessment tasks (mean=193.80) than when completing fact tasks (mean=128). For number of clicks ($F(1,57) = 585.126$, $p<0.001$); number of additional pages ($F(1,57) = 382.867$, $p<0.001$); number of correct tasks ($F(1,57) = 598.162$, $p<0.001$), participants completed more fact tasks (mean= 0.65) than Assessment tasks (mean= 0.49). For number of abandon tasks ($F(1,57) = 141.858$, $p<0.001$); number of partly completed tasks ($F(1,57) = 46.658$, $p<0.001$) and number of incorrect solutions ($F(1,57) = 88.316$, $p<0.001$).

| | **Silent working** | | **Classic think-aloud** | | **Explicit Instruction** | |
|---|---|---|---|---|---|---|
| | Fact | Assessment | Fact | Assessment | Fact | Assessment |
| | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| Time on tasks* | 139.80 (38.39) | 194.40 (34.73) | 123 (40.06) | 192.60 (48.85) | 121.20 (35.58) | 194.40 (53.46) |
| Number of clicks | 13.11 (5.91) | 18.19 (7.02) | 11.59 (8.23) | 20.35 (5.37) | 10.68 (2.83) | 19.10 (6.09) |
| Additional pages | 2.08 (1.10) | 2.44 (1.30) | 1.90 (1.00) | 3.06 (1.46) | 1.97 (1.06) | 2.80 (1.18) |
| Correct tasks* | 0.56 (0.24) | 0.35 (0.23) | 0.72 (0.23) | 0.59 (0.25) | 0.68 (0.19) | 0.53 (0.22) |
| Abandon tasks | 0.20 (0.20) | 0.37 (0.21) | 0.13 (0.14) | 0.28 (0.22) | 0.14 (0.16) | 0.23 (0.18) |
| Partly completed | 0.03 (0.07) | 0.17 (0.22) | 0.01 (0.04) | 0.15 (0.14) | 0.01 (0.04) | 0.17 (0.14) |
| Incorrect solutions | 0.19 (0.12) | 0.14 (0.20) | 0.14 (0.14) | 0.09 (0.12) | 0.15 (0.14) | 0.13 (0.13) |

Table 5.4 Performance data for classic; explicit and silent condition

*Table 5.4 Performance data for classic; explicit and silent condition*

### 5.6.2 Mental Workload

Table 5.5 summarises the TLX subscales results of the three conditions: classic think-aloud, explicit instruction and silent conditions together with task types: fact and assessment tasks respectively. The significant results are marked with asterisk.

Findings indicated that explicit instruction led to high mental workload in terms of performance over classic think-aloud and silent condition, however, classic think-aloud and Silent led to high mental workload in terms of effort. The TLX subjective measures results for mental workload are consistent with the performance measures. Specifically, mental workload correlated with the amount of effort put in for assessment tasks for all three conditions. Although, effort was more pronounce on the silent condition during assessment task, suggesting that participants may have put in extra effort since they were carrying out task performance in silent.

Table 5.5 summarises the TLX workload for classic, explicit and silent conditions

| | **Classic Think Aloud** | **Explicit Instruction** | **Silent Working** |
|---|---|---|---|
| | Mean (SD) | Mean (SD) | Mean (SD) |
| **Overall mean** | 47.01 (16.43) | 45.98 (18.05) | 49.20 (19.96) |
| **Mental** | 47.25 (17.92) | 41.60 (19.24) | 56.22 (22.59) |
| **Temporal** | 37.67 (21.45) | 34.55 (17.65) | 40.30 (21.95) |
| **Performance** | 71.75 (15.79) | 74.65 (14.30) | 63.35 (18.00) |
| **Effort** | 50.55 (15.90) | 43.75 (19.03) | 52.70 (20.11) |
| **Frustration** | 27.82 (11.49) | 35.35 (20.05) | 33.42 (17.15) |

*Table 5.5 summarises the TLX workload for classic, explicit and silent conditions*

Table 5.6 summarises the TLX subscales results of the three conditions: classic think-aloud, explicit instruction and silent conditions together with task types: fact and assessment tasks.

In terms of the impact of task on participants' workload, participants on the explicit condition reported higher level of frustration when completing assessment task mean (42.50) when compared to the classic mean (30.60) and silent condition (37.90).

Table 5.6 TLX workload for classic, explicit and silent condition with fact and assessment task.

Table 5.6 Scale: 0 (very low) to 100 (very high)

| | **Classic Think Aloud** | | **Explicit Instruction** | | **Silent Working** | |
|---|---|---|---|---|---|---|
| | Fact tasks | Assessment | Fact tasks | Assessment | Fact tasks | Assessment |
| | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Mental** | 38.40 (19.63) | 56.10 (19.63) | 33.05 (18.09) | 50.15 (24.40) | 47.05 (24.26) | 65.40 (24.2) |
| **Temporal** | 37.55 (23.04) | 37.80 (21.04) | 31.00 (18.09) | 38.10 (21.12) | 40.90 (22.68) | 39.70 (24.15) |
| **Performance** | 75.13 (25.43) | 64.65 (19.45) | 79.65 (14.35) | 69.65 (16.86) | 70.90 (19.72) | 55.80 (22.91) |
| **Effort** | 45.20 (21.47) | 55.90 (14.27) | 36.65 (18.53) | 50.58 (22.31) | 44.75 (20.30) | 60.65 (22.82) |
| **Frustration*** | 25.05 (14.12) | 30.60 (13.03) | 28.20 (18.31) | 42.50 (23.56) | 28.95 (17.66) | 37.90 (23.91) |

*Table 5.6 TLX workload for classic, explicit and silent condition with fact and assessment task.*

### 5.6.3 Participants' perceptions

After the task completion, participants were asked about their think-aloud content to find out if participants were self-monitoring. Hence, participants were asked to indicate the extent to which they agreed or disagreed with several statements during think-aloud.

Table 5.7 presents how participants rated their task performance and test experience. A Mann Whitney U test revealed no significant difference between the two think-aloud methods ($U=110$, $p<0.001$). Participants on the explicit condition (mean =15.67) reported higher agreement in terms of the number of successfully completed tasks compare to those on the classic condition (mean =15.33).

*Table 5.7: Participants self-reported questionnaire for classic and explicit condition Finding higher agreement*

| | **Classic** **Mean (SD)** | **Explicit instruction Mean (SD)** |
|---|---|---|
| I was able to concentrate during task performance | 4.35 (0.67) | 4.55 (0.68) |
| Thinking aloud interfered with my performance during the tasks | 2.85 (1.38) | 2.9 (1.55) |
| I was worried about talking too long on those tasks I found difficult | 3.25 (1.44) | 3.0 (1.33) |
| The presence of the test facilitator made you feel uncomfortable | 2.1 (1.48) | 2.0 (1.41) |
| I persisted with tasks for longer than I would normally do so in real work use | 2.7 (1.55) | 2.9 (1.29) |
| The things I said during my think-aloud reflected all of my thoughts about the tasks | 3.6 (1.14) | 4.0 (0.97) |
| I withheld some information from my think-aloud | 2.8 (1.67) | 2.4 (1.31) |
| I was concerned about giving up early on those tasks I found difficult. | 3.05 (1.53) | 3.1 (1.33) |
| Were any of these issues a factor in you giving up? | | |
| Burdon | 0.05 (0.22) | 0.1 (0.30) |
| Time factor | 0.35 (0.48) | 0.3 (0.47) |
| Frustration | 0.35 (0.48) | 0.5 (0.51) |
| Unachievable task | 0.3 (0.47) | 0.3 (0.47 |
| Others | 0.05 (0.22) | 0.1 (0.30) |
| I prefer to work in silence | 3.6 (1.46) | 3.1 (1.29) |
| I feel satisfied with the number of successfully completed tasks | 3.7 (0.97) | 4.0 (1.25) |

Scale: 5 strongly agree to 1 strongly disagree *: a significant p<0.5 obtain

When ask if there were able to concentrate during task performance, participant on the classic condition reported that “the environment is conducive enough to complete task. While one participant on the explicit condition reported that “yes I was able to as the website showed menu for metro, bus etc” and another said “I like efficient and seem like tasks time is precious to me”

When ask to comment on think-aloud interference, two participants, here after “P”, on the classic condition reported that “I think think-aloud would distract you more” [P11] and the other reported that “I am a very quiet person and would normally concentrate more if quiet” [P19]. Although participants on the explicit condition did not comment on their think-aloud interference, one plausible explanation to this could be linked to social desirability as some participants wouldn’t want to be perceived negatively so they over report good behaviour or under report undesirable behaviour.

When ask if there were worried about taking too long on tasks, they found difficult one participant on the classic condition reported that “I wanted to concentrate to find the

information I was looking for" [P12] and another reported that "I wasn't worried really" [P7]. While one explicit participant reported that "I was cautious about time record for a task". When ask if the test facilitator made participant feel uncomfortable, one participant on the classic think-aloud reported "Not at all" [P18] and another reported that "because at the moment I was feeling like nervous due to the facilitator's presence" [P6].

When ask if they withheld some information during think-aloud, one participant on the classic condition reported that "during frustration I had to stop myself from swearing aloud" [P9].

Overall, participant on the classic condition tends to verbalise their thought during tasks performance, while participants on the explicit instruction tends to focus on what likes, dislikes or find confusing about the website and a bit worried about time factor.

Table 5.8: Participants self-reported questionnaire for silent condition

| | **Silent**<br><br>**Mean (SD)** |
|---|---|
| I was worried about talking too long on those tasks I found difficult | 3.05 (1.31) |
| The presence of the test facilitator made you feel uncomfortable | 1.6 (0.99) |
| I persisted with tasks for longer than I would normally do in real world use | 2.65 (1.26) |
| I was concerned about giving up early on those tasks I found difficult. | 2.45 (1.14) |
| Were any of these issues a factor in you giving up? | |
| Burdon | 0.1 (0.30) |
| Time factor | 0.3 (0.47) |
| Frustration | 0.1 (0.30) |
| Unachievable task | 0.3 (0.47) |
| Others | 0.15 (0.36) |
| I feel satisfied with the number of successfully completed tasks | 3.75 (0.96) |

Scale: 5 strongly agree to 1 strongly disagree *: a significant p<0.5 obtain

### 5.6.4 Verbal Utterance

Table 5.9 presents the data for verbal utterances for both the concurrent think-aloud and the explicit instruction. The total number of utterances and the number of utterances made in each category for the think-aloud method: classic think-aloud and explicit instruction.

On the classic think-aloud condition participants produce more utterances (2478) compared to explicit instruction participants (2362). On the classic condition the most dominant utterance category was "action description" (37.5%) and "action explanation" (17.8%). Whereas for the explicit condition the most dominant was "action description" (34.6%) and causal explanation" (23.8%). Although, the explicit instruction did produce more verbal utterances in categories that usability practitioners expected and find relevant such as utterance category in user “experience” and “expectations”.

Table 5.9 utterance categories of classic think-aloud and explicit instruction

| **Utterance categories** | **classic think-aloud** | | **explicit instruction** | |
|---|---|---|---|---|
| | **Sum** | **Mean (SD)** | **Sum** | **Mean (SD)** |
| **Total** | **2478** | **123.9 (3.244)** | **2362** | **118.1 (5.881)** |
| Action Description | 931 | 46.55 (1.833) | 818 | 40.9 (0.324) |
| Action Explanation | 442 | 22.1 (5.831) | 370 | 18.5 (0.324) |
| Causal Explanation | 434 | 21.7 (1.007) | 561 | 28.05 (0.324) |
| Domain Knowledge | 0 | 0 (0) | 1 | 0.05 (0.003) |
| Expectation | 16 | 0.8 (0.056) | 25 | 1.25 (0.003) |
| Help Request | 15 | 0.75 (0.006) | 23 | 1.15 (0.003) |
| Impact | 33 | 1.65 (0.016) | 39 | 1.95 (0.003) |
| Problem Indication | 197 | 9.85 (0.407) | 189 | 9.45 (0.324) |
| Reading | 225 | 11.25 (0.324) | 165 | 8.25 (0.858) |
| Recollection | 6 | 0.3 (0.065) | 0 | 0 (0) |
| Recommendation | 1 | 0.05 (0.015) | 3 | 0.15 (0.006) |
| Result Evaluation | 153 | 7.65 (0.649) | 133 | 6.65 (0.324) |
| Tasks Confusion | 12 | 0.6 (0.032) | 3 | 0.15 (0.013) |
| User Experience | 13 | 0.65 (0.003) | 32 | 1.6 (0.141) |

Figure 5 Utterance categories for classic think-aloud & explicit instruction. In terms of the utterances that test facilitators desires, the explicit condition produce more utterance category for “user experience” accounting for (1.35%) and “expectation” (1%) whereas, for the classic

condition utterance category for “user experience” accounting for approximately (0.6%) and “expectation” account for approximately (0.7%).

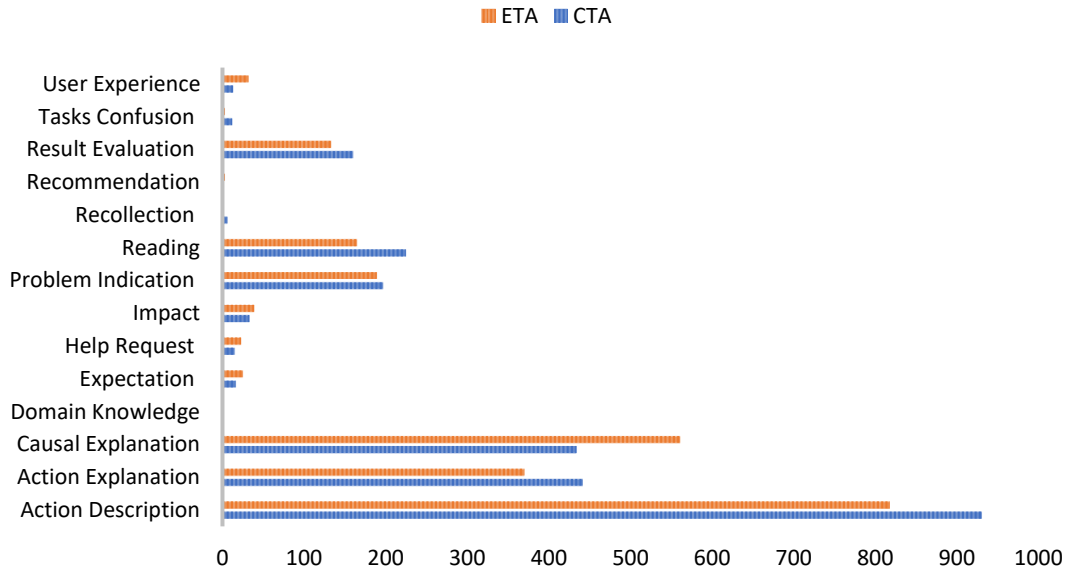



*Figure 5: Utterance categorise for classic and explicit instructions*

Table 5.10 shows the number of utterances for fact and assessment both classic think-aloud and explicit instructions. On the classic condition a total number of 1061 was produce for fact task and 1417 for assessment task. While for the explicit condition a total number of 1003 was produce for fact task and 1359 for assessment task.

Table 5.10 utterance categories of classic think-aloud and explicit instruction by task-types: fact and assessment

| Utterance categories | Classic think-aloud | | Explicit Instruction | |
|---|---|---|---|---|
| | Fact task | Assessment task | Fact task | Assessment task |
| **Total** | **1061** | **1417** | **1003** | **1359** |
| **Action Description** | 400 | 531 | 378 | 440 |
| **Action Explanation** | 197 | 245 | 150 | 220 |
| **Causal Explanation** | 160 | 274 | 180 | 381 |
| **Domain Knowledge** | 0 | 0 | 1 | 0 |
| **Expectation** | 8 | 8 | 12 | 13 |
| **Help Request** | 6 | 9 | 10 | 13 |
| **Impact** | 9 | 24 | 17 | 22 |
| **Problem Indication** | 70 | 127 | 68 | 121 |
| **Reading** | 118 | 107 | 105 | 60 |
| **Recollection** | 2 | 4 | 0 | 0 |
| **Recommendation** | 0 | 1 | 1 | 2 |
| **Result Evaluation** | 80 | 73 | 70 | 63 |
| **Tasks Confusion** | 3 | 9 | 0 | 3 |
| **User Experience** | 8 | 5 | 11 | 21 |

On the classic condition the most dominant utterance category was "action description" (37.7%) and "action explanation" (18.6%) for fact task, then for assessment task it was "action description" (35.3%) and "causal explanation" (19.3%). whereas, for the explicit condition the most dominant was "action description" (37.7%) and causal explanation" (17.9%) for fact task, and for assessment task it was "action description" (32.4%) and "causal explanation" (28%).

Figure 3 shows the total number of utterances categories produced for both fact and assessment task on the classic think-aloud and the explicit condition. As shown on the diagram the dominate utterance category was “action description” for assessment task on the classic think-aloud, followed by “action description” for assessment task on the explicit condition. This is followed by “action description” for fact task on the classic and then the explicit condition.

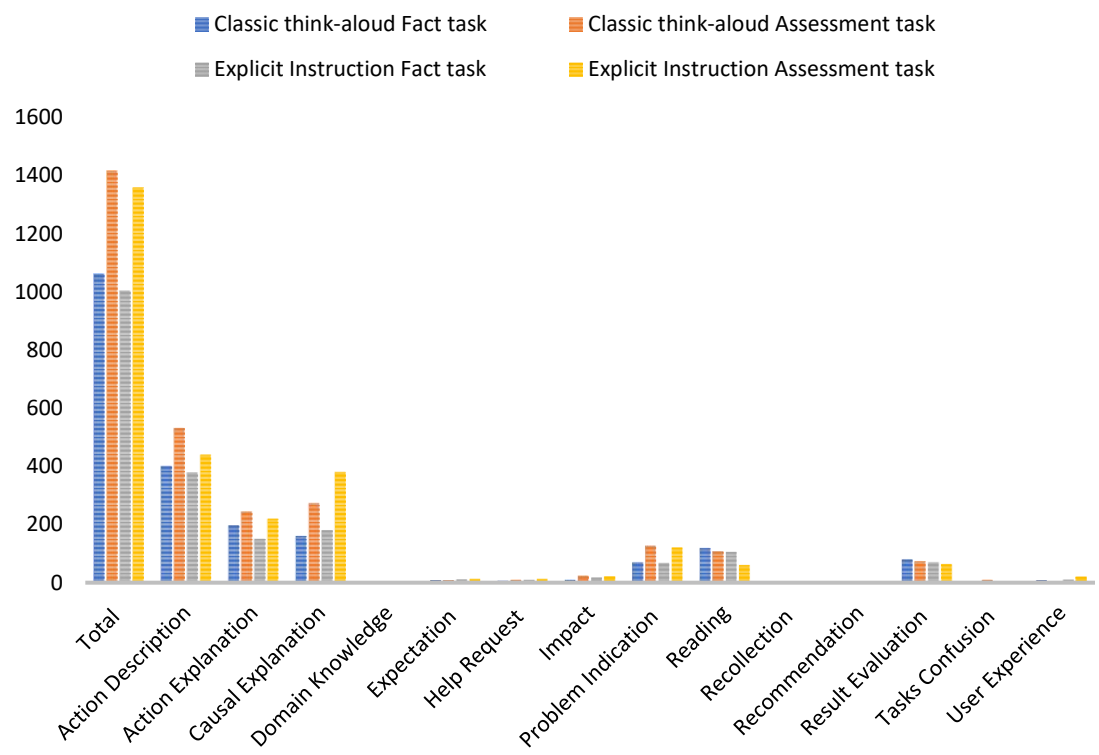



*Figure 6 Utterance categories of classic think-aloud explicit instructions with fact and assessment tasks*

## 5.7 DISCUSSION

### 5.7.1 Performance

Based on the findings of this study we reject H1: participants on the explicit condition are more likely to perform better in terms of task success, mouse clicks and number of additional pages on both fact and assessment task. When completing assessment tasks. explicit instruction participants spend more time than classic think-aloud participants and classic think-aloud participants spend more time than participants on the silent condition than when completing fact tasks. Although, classic think-aloud slightly outperformed explicit instruction, there were no further interaction between task-type and think-aloud indicating non-reactivity of the classic think-aloud for this study.

Findings from this Study corresponds to earlier findings by Hertzum, Hansen and Andersen, (2009) confirming that the classic think-aloud has little impact on participants behaviour and mental workload, except for task elongations. Also, findings from Dickson, McLennan and Omodei, (2000) in which they investigated the effects of concurrent verbalisation on a time-critical, dynamic decision-making task using a computer simulation to fight a forest fire. Findings indicated that, participants who verbalised reasons for their actions performed worse than participants who did not verbalise and that the performance of participants who thought aloud was intermediate between the two other conditions and no different from any of them. Thus, implying that thinking aloud does not have a direct impact on time constraints, however, the study did not provide evidence to verify if participants were under time pressure.

In terms of numbers of clicks, there was a significant main effect of task as participants made more mouse clicks when completing assessment tasks than fact tasks, although there was no link between task and think aloud. One might suggest that it could be linked to the level of task

difficulty between fact and assessment task, as participants made more effort and engaged in more clicking and scrolling activities to obtain information on the web pages. This findings accords with Hertzum, Hansen and Andersen, (2009) which indicated that participants on the classic condition made marginally more clicks than participants on the silent condition. And for the relaxed think-aloud condition participants made more clicks and scrolling activities when compared with the silent condition.

Also, participants navigated more pages and carry out more scrolling activity during assessment tasks than when completing fact task and this was more pronounced on classic think-aloud participants than explicit instruction and participants on the silent condition. This suggest that participants engaged in more link traversals, potentially seeking different paths to solve the underlying task (McDonald, McGarry and Willis, 2013b).

In terms of successfully completed tasks, result indicated that participants completed more fact tasks than assessment tasks and there was no further interaction between task and think-aloud. In terms of task success, task performance was not affected by the extra workload to simultaneously think-aloud during task performance both in the classic think-aloud and explicit instruction conditions. A possible explanation of this finding is task difficulty, due to different task-type. Proponent of task difficulty McDonald, McGarry and Willis, (2013),investigated the relationship between think-aloud instructions, task difficulty and performance. Their findings indicated that for more difficult tasks, participants in the classic condition completed fewer tasks successfully and engaged in more clicking activities than participants in the explicit instructions condition. However, for low difficulty tasks, there was no difference in task success between the two conditions.

Also, regarding the use of instructions findings from McDonald, Edwards and Zhao, (2012) indicated that there were no difference between Explicit instruction and the classic think-aloud

method in terms of task performance, except for the classic method leading to an increased mental workload and explanatory utterances. In addition, (Krahmer and Ummelen, 2004) reported on an exploratory experiment comparing two think-aloud approaches: the classic think-aloud and a variant of relaxed thinking-aloud proposed by Boren and Ramey (2000), findings indicated that, think-aloud during task performance has no impact on the type of usability evaluation method that was adopted, although, the relaxed think-aloud method led to more correctly solved task when compared to the classic method. A plausible explanation to the differences in our result and that of Krahmer and Ummelen, (2004) could be task abandonment, suggesting the simplicity of fact task compare to assessment task which were more difficult.

Focusing on behaviour and mental workload, Hertzum, Hansen and Andersen, (2009) investigated whether and how think-aloud in the classic or relaxed way influences people's behaviour. They found that the classic think-aloud has no significant impact on participant's behaviour apart from tasks elongation. Whereas, during relaxed think-aloud participants spent more time on task, engaged in more link traversals and experienced higher mental workload. These findings were somewhat different from those identified by Van den Haak and de Jong, (2003) which explore two methods of usability testing: concurrent versus retrospective think-aloud methods. Their findings indicated that, participants in the concurrent think-aloud performed less successful compared to those working in silent and the retrospective condition.

Although, a previous study conducted by Wright and Converse, (1992) which investigated the impact of concurrent verbalisation on task performance indicated that, participant verbalisation at level 3 during task performance led to fewer errors when compared with silent working. They linked this high success rate in task performance to a better understanding of task which is achieved by prompts, requesting participants to explain the reasons for their behaviours.

The most plausible explanation for this disagreements between our result and that of Wright and Converse, (1992) is that, participants on the explicit instruction conditions where not prompt to give reasons for their behaviour rather, an explicit instruction was given at the beginning of the test. In terms of number of incorrect solutions and abandon tasks, findings indicated that participants produce more incorrect solution during explicit instruction than on classic think-aloud and the silent condition. Also, more incorrect solution on fact task than on assessment task which were mostly abandon, there were no further evidence linking this to think-aloud condition.

One plausible explanation to the reason why there were more incorrect task for fact task could be because participants did not interpret the solution on the website correctly or did not fully understanding the task. For example, when participants were asked to find out the cost of a student metro season ticket for one month, which covers all zones. Most participants wrote and announced the annual cost of the ticket instead of four weeks cost as their final answer. Hence, since fact task entails looking for information that is explicitly available on the website, it is more likely they make the wrong choice and assume they have successfully completed the task. Unlike the assessment task which were partly completed and more likely abandon due to task difficult and frustration in which participants gather information and based on this information formed an opinion.

Also, the author partly accepted H2: The use of explicit instruction with fact and assessment task might lead to an increase in participant mental workload. Findings indicated that the explicit condition led to high mental workload in terms of performance over the classic condition, although, the classic did lead to high mental workload in terms of effort.

### 5.7.2 Participant Utterances

H3 was rejected based on experimental results. The use of explicit instruction with fact and assessment may lead to an increase in participants verbalisation compared to the classic condition. Findings from the Study indicated that, overall classic think-aloud participants produce more utterances (2478) than explicit instruction participants (2362). A plausible explanation to this could be the different types of instructions associated with the two conditions, see section table 5.2 for details of instruction for the two conditions. For explicit instruction condition utterances were more pronounced for action description, casual explanation, action explanation and reading. While for classic think-aloud utterances were more pronounce for action description, action explanation and causal explanation. Although, participant in the explicit instruction condition produced more utterances in terms of “User Experience” compared to classic think-aloud condition. See table 5.10 for details.

While for the classic think-aloud participants based on the instruction give “I want you to say out loud everything that you say to yourself in silent” ended up giving recommendation and also verbalised about the user experience of the site even though they were not instructed to do so, a plausible explanation could be because they were told before the test that they were going to evaluate the user-friendliness of the website.

H3b was accepted based on experimental results. Utterance category, user experience (32) and expectations (25) for explicit instruction condition compared to classic think-aloud condition which has user experience (13) and expectations (16) for as the total number of utterances produced by participants. One plausible explanation to this is due to the fact that participants on the explicit instruction condition focus on the instruction which may result to them being sensitive to issues relating to the user-friendliness of the tested product.

## 5.8 Limitation and future work

The main limitation of this study is that the author functioned as both the test facilitator and the data coder. Ideally, different individuals would have performed the coding activities. To mitigate the potential bias this introduced the following measures were taken: (i) there was a delay of three weeks between data collection and transcription and a further five weeks for subsequent qualitative analysis; (ii) the second author crossed checked all qualitative data without knowledge of which data belongs to a particular think-aloud condition.

Future research could further delve into the implications of explicit instruction within usability testing by specifically focusing on the timing and nature of the instructions provided. An important aspect that could be investigated is how the timing of the instructions impacts user performance. For instance, instructions provided at the beginning, during, or after a task may lead to different outcomes. By meticulously recording and analysing these timelines, researchers could determine optimal instructional intervals that promote both task completion and high levels of verbalisation.

In addition to timing, the explicitness of the instruction might also influence the results. While some instructions could be broad or high-level, others might be specific and detailed. To understand these effects, studies could compare user performance under various explicit instructional conditions. A potential way to ensure that the instructions are consistent across tests is by utilising video demonstrations. This not only eliminates possible variations in verbal instruction but also provides a visual aid which may prove beneficial for certain types of tasks or individuals.

Moreover, the utilisation of video/timeline analysis would undoubtedly offer a robust method for collecting data. Video analysis, for example, would allow researchers to observe user reactions to different types of explicit instructions, and to note any correlations between these reactions and performance outcomes. Timestamped markers could help identify when certain

actions or difficulties occur, providing a more granular understanding of the user's experience and reaction to the instructions given.

This two-pronged approach, focusing on both the timing and the explicitness of instruction, could significantly enhance our understanding of the instructional parameters that affect usability testing outcomes. It can open up new ways to optimise instructions for better user interaction and performance, which would be a valuable contribution to the field.

## 5.9 Summary

This study explores an important and active area of research, the use of explicit instruction within usability testing, focusing on issues relating to the working habits of practitioners. Findings from this Study contradicts previous studies on the use of the explicit instruction. The classic think-aloud did provide useful indication in terms of its impact on task performance: classic think-aloud participants completed more tasks successfully than participants in the explicit instruction condition and shows no indication of reactivity.

Although, the explicit instruction did produce more verbal utterances in categories that usability practitioners expected and find relevant. However, explicit instruction participants completed fewer tasks successfully and the explicit instruction led to high mental workload in terms of performance over the classic think-aloud as indicated by the TLX subscales results.

One important finding from this study shows that the explicit instruction produced less verbalisation when compared to the classic think-aloud which led to an increase in relevant explanatory utterances. Given that the explicit instruction explicit request participants to verbalise "things they like, dislike and find confusing" about site one may expect an increase in the total number of relevant explanatory utterances. However, the reverse was the case as classic think-aloud condition produced more verbal utterances than explicit instruction condition.

This could be a case of verbal overshadowing of process and criterion, where the formal refers to a shift in processing caused by verbalisation and the latter refers to the possibility that verbalisation led to a reliance on more controlled choosing, suggesting that the use of explicit instruction make users focus on things they like, dislike or finding confusing about the tested product in order to act in accordance with the explicit instructions given by the test facilitator.

# CHAPTER SIX

## 6.0 PRACTITIONERS' USE OF CONCURRENT THINK-ALOUD: PRACTISES AND CHALLENGES

### 6.1 Overview

This chapter presents the third study of this PhD research. This study explores how UX practitioners use the concurrent think-aloud method within usability testing in the industry, practitioners' views on reactivity and the challenges they faced with using the think-aloud protocol in the industry. Findings obtained from the study were discussed, and the chapter concludes and presents recommendations and future research.

### 6.2 Motivation

This PhD research started by providing evidence to support previous research that the classic think-aloud is not reactive. See the first empirical investigation presented in this thesis (Chapter four). The second study looked at differences in task performance. The result shows no evidence of reactivity and there were no significant main effects or interactions between task performance and think-aloud, these results warrant further investigation. This forms the premise for a third study, which initially was planned to investigate whether the presence of a test facilitator has an impact on participant performance to cause reactivity. However, the study coincided with the pandemic (see chapter 1, p13 for details). Hence, the author was not able to do lab work. While the pandemic meant the author could not proceed with the initial lab work, it was a suitable time for the author to explore other factors and this meant looking at practitioners' use of the concurrent think-aloud protocol. This ties in with the two studies in this thesis as it explores practitioners' views on reactivity. The study focuses on the following themes: (i) characteristics of the think-aloud test (ii) nature of tasks practitioners were using and (iii) practitioners' views on reactivity and interacting with participants during think-aloud sessions. Understanding the practices and challenges of using the think-aloud method in the industry is critical as UX approaches improves.

The ability to reflect on and learn from the practices of the UX community is aided by this awareness. Academic researchers and tutors must also be aware of the implementation and challenges that arise when using the think-aloud protocol within usability tests in the industry.

## 6.3 Study Aims and Objectives

This study examines practitioners' experiences, views on reactivity and challenges when using the think-aloud method within usability testing to make an informed decision when using the think-aloud protocol within usability test and to explore methods to address these challenges.

### 6.3.1 Research Questions

**RQ3 – What are the practices and challenges of using the think-aloud protocol in the industry?**

(i) How do UX practitioners use the think-aloud method within usability testing?

(ii) What is the nature of tasks practitioners uses?

(iii) What are practitioners' views on reactivity?

## 6.4 METHODOLOGY

In this section, the author discusses the choice of method, identification of interview questions, and implementation of the adopted approach, which is used to gather qualitative data from usability professionals in commercial practice or industry.

### 6.4.1 Choice of Method

Evidence from usability research on how think-aloud methods are being used in practice has been shaped by previous research (Fan et el., 2020; McDonald et al., 2012; Boren and

Ramey, 2000; Nørgaard, & Hornbæk, 2008 and Shi, 2008). Others have helped shape usability research and have provided a thorough account of how practitioners should use the concurrent think-aloud protocol (McDonald, Zhao & Edwards, 2016; Zhao, McDonald & Edwards, 2014 and Zhao & McDonald 2010).

This study investigates the characteristics of a usability test to understand the challenges UX practitioners face in the industry within the United Kingdom when implementing the think-aloud protocol in various practical contexts and explore better methods to address these challenges.

The author adopts a semi-structured interview research method; In a semi-structured interview, prepared questions are asked consistently and systematically, guided by identified themes. The focus is on the interview guide, which includes a series of broad themes to be covered throughout the interview to assist lead the conversation toward the topics and concerns that the interviewers are interested in learning about (Kvale and Brinkmann, 2009).

The rationale for using semi-structured interviews is that they reveal crucial and frequently overlooked human and organisational behaviour aspects. They allow interviewees to respond in their own words, in the way they think and use language.

### 6.4.2 Interview Design

In this section, the author discusses the interview process, including designing the interview questions, recruiting participants, running a pilot study, and conducting the interview.

#### 6.4.2.1 The Process of Designing and Developing the Interview Questions

The process of designing and developing the interview questions was informed by the usability literature on think-aloud studies. The author wanted to examine how practitioners followed the established guidelines given by Ericsson and Simon when implementing the think-aloud techniques during usability testing.

Hence, interview questions were derived from three phases of usability testing: (a) the planning phase, which includes the study goals, building rapport, participant, test environment, method implementation and tasks. For instance, RQ3 (i) how do UX practitioners use the think-aloud method within usability testing? It was derived from the planning phase of a usability test and usability literature such as Fan et al. (2020). (b) testing phase includes the instructions, tasks and approach used by practitioners to interact and request verbalisation from participants.

For instance, the RQ3(ii) What is the nature of tasks practitioners uses? It was derived from the testing phase and usability literature such as McDonald and Petrie (2013). And RQ3(iii) practitioners' views on reactivity were derived from the testing phase and usability literature such as Hertzum et al. (2015). (c) analysis phase: usability problem measures and report writing. For instance, in RQ3(iv) how do practitioners analyse the data obtained from a usability test? This was derived from the usability testing analysis phase and studies such as McDonald et al. (2012).

#### 6.4.2.2 Running a Pilot Interview

The author iterated through the interview questions to refine them, eliminate ambiguities and ensure that the processes throughout the interview were straightforward. Ensuring that the theme of interest had been addressed and that the interview duration would be completed within 30 minutes. The interview was then piloted with 4 participants within the usability field.

This pilot test provides insight into the use of think-aloud and comments on the clarity of the questions being asked and the interview structure. Some analyses indicated misinterpretations of questions, while others suggested concerns that people thought were significant but had been overlooked. The feedback was utilised to examine the interview questions and identify areas that required improvement, such as rephrasing questions for better clarity, reducing the number of questions asked and focusing on the study's primary focus of reactivity.

For example, interview questions such as “what is your approach to usability test” and do you see it as a necessary step in software development? It was reversed to “Do you see think-aloud as an essential part of usability testing? This shows more clarity based on the feedback received from the pilot study.

The final interview questions consisted of 24 questions which focus on how UX practitioners use the think-aloud method, practitioners' views on reactivity and how practitioners analyse the data obtained from a usability test. The interview questions show more clarity and are in accordance with the research questions. See appendix C4 for details of interview questions.

#### 6.4.2.3 Interview Procedure: How the Interview was Conducted

Consent forms and screening questionnaire was designed using Qualtrics (see appendix C2). Practitioners (representative participants) were contacted via LinkedIn. The interview process was conducted online using Microsoft Teams. The author started the interview sessions by welcoming the interviewee, explaining what was going to happen, and stating the interview's purpose and duration. Interviewees were told that the interviewer would be taking notes during the interview. Specifically, the following instructions were given to all interviewees: "When we start talking, I am interested in your personal experience; there are no right or wrong answers.

So, whatever comes to mind is fine. You can withdraw from the interview at any time if you don't want to continue. So please, do you have any questions before we start?" An ice-breaking question followed: do you see think-aloud as an essential part of usability testing? At the end of the interview, participants were debriefed and reminded of how their data would be used. And the findings will be summarised and published on LinkedIn.

The interviewee was also allowed to ask any questions concerning the study. At the end of the interview, they were thanked for voluntarily participating in the study.

#### 6.4.2.4 The process of Recruiting Participants

The study participants were UK-based UX practitioners. The importance of selecting the appropriate candidates for the interview was a top priority for the study. A criterion-based sampling (UX practitioners conducting usability testing on users) was implemented to obtain qualified candidates willing to openly and honestly share their experience to provide the most credible information to the study.

Hence the recruitment process was done on LinkedIn, which has a pool of representative participants for the study. The author created a post on LinkedIn, signposting the study's objective with a link to a screening form. Only participants that have conducted usability

testing with users as part of their job were further contacted to provide their email addresses to progress with booking an interview session by scheduling a date and time for the interview. Only two participants were told that they did not meet the interview criteria clearly stated on the post (practitioner that conducts usability testing using the think-aloud techniques).

Qualified participants were sent a link to the participant information sheet that details the study's objective and how the data obtained from the interview will be used.

### 6.4.3 How Data Was Transcribed

An interview study requires the data obtained to be transcribed to aid analysis. The study was done on Teams during the pandemic; hence, the author made use of initial text-to-speech software and then replay the recording to check for transcription errors. Hence, the rationale for using an online audio-to-text automatic transcription tool (Temi: https://www.temi.com/).

After transcribing each session automatically, the author reviewed each interview session by listening to the audio recording and making corrections manually where the online tool failed to recognise certain words used by both the interviewer and the interviewees. This ensures that the data have been transcribed to an appropriate level of detail and the transcripts have been checked against the recordings for accuracy.

## 6.5 RESULTS ANALYSIS

### 6.5.1 Participant's Profile

Qualification: The author asks participants thereafter practitioners about their level of qualification and most of the participants reported their level of qualification as a bachelor's degree (68%); followed by a master's degree which was (18%) and PhD (14%).

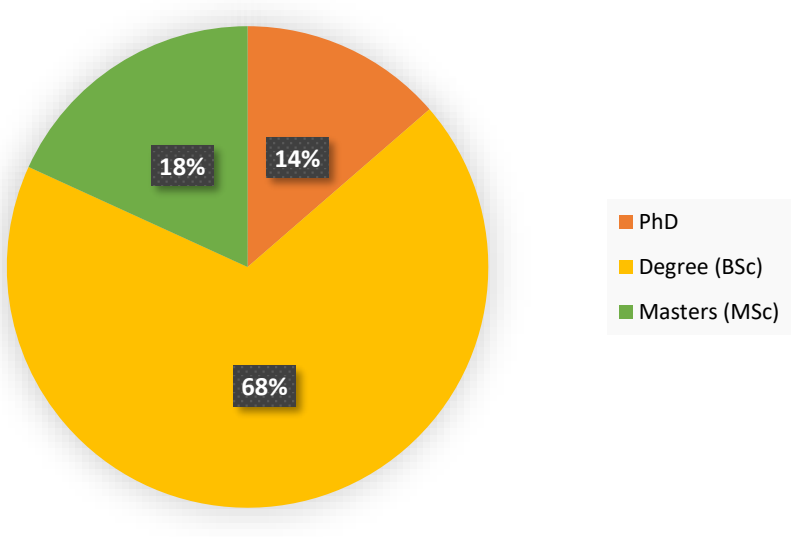



*Figure 7: Participants Educational Background (n=22)*

**Work Experience:** Questions in the questionnaire include the number of years participants had worked in the UX industry/Usability testing fields. Participants reported their work experience within the UX industry. Most of the participants reported to have above 5 years of experience (50%); followed by participants with exactly 5 years' experience (23%); followed by participants with above 15 years' experience (18%) and 20 years of experience (9%).

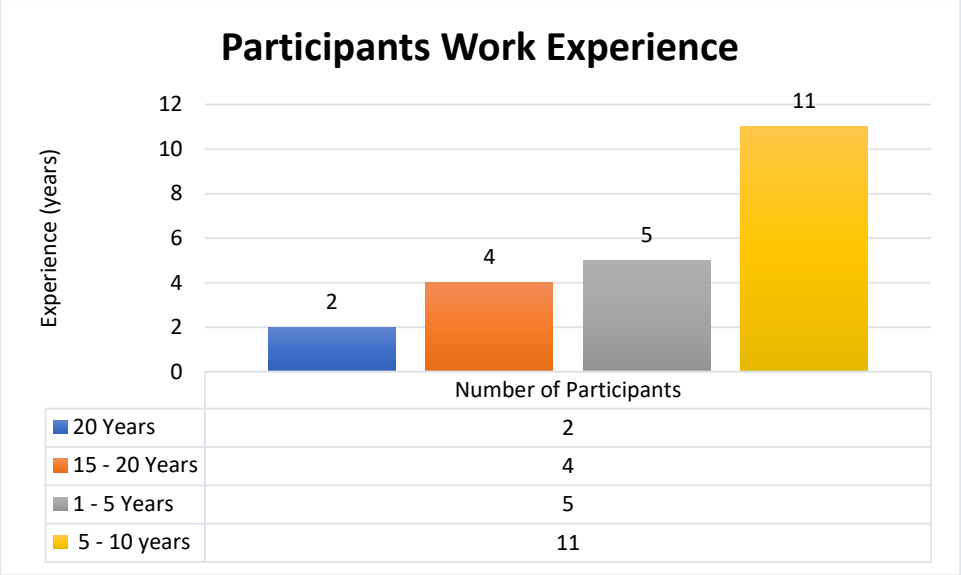



*Figure 7: Participants work experience*

**Job Role:** Participants reported their current job title. Most participants reported their current job title as UX Researcher (86%), followed by UX Manager (9%) and UX Lead (5%).

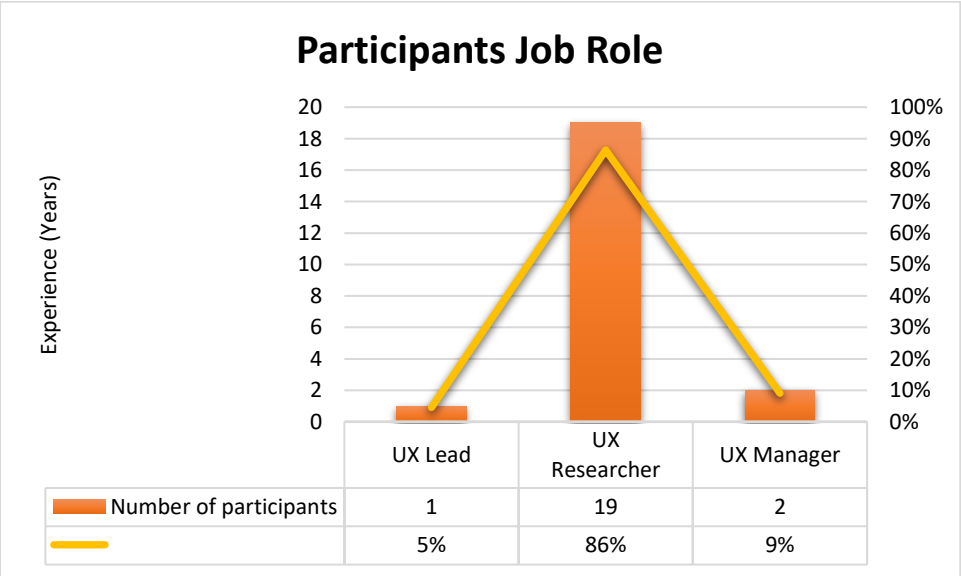



*Figure 8: Participants job roles*

**Organisation:** In terms of organisation participants' profiles covered a wide range of industrial fields theses including UX consulting, marketing, software development, gaming, banking, healthcare, housing, and Telecommunication. See full details of participant profile in appendix C5.

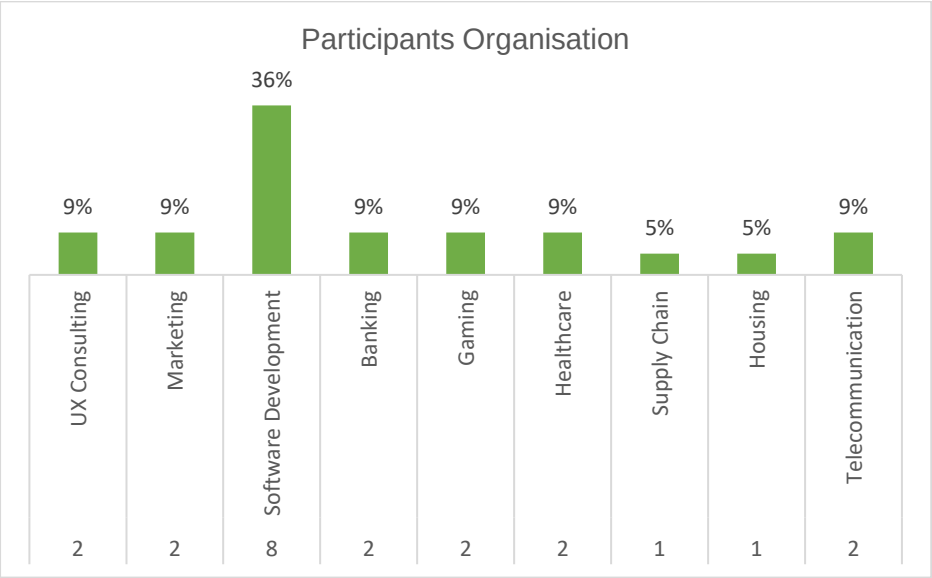



*Figure 9: Participants' organisation*

## 6.6 Data Analysis

This section details the findings from the study, which is based on an online interview with 22 UX practitioners. The interview revealed several themes, but the most important ones which can be related to the study's research questions are included in table 6 below. In this section, the author analyses the themes in detail. Some direct quotations from the raw data were utilised to show clarity and complement descriptions to demonstrate the authenticity of the interviews, which will provide a clear understanding of the relevance of the underlying issues from the interviewee's perspective, allowing for a clearer interpretation. Table 6 shows the main themes and their associated sub-themes.

*Table 6: main themes and their associated sub-themes*

| **Emerging Themes** | **Sub-themes** | **Transcript** |
|---|---|---|
| Think-aloud Usefulness | Expeditious<br>Task solving strategy<br>Insights<br>User experience<br>User preference<br>Rapport | *"it's the thing that will give us insights into people's, um, expectations"* |
| Instructions used during think-aloud | Explicit instructions<br>Neutral instructions<br>Reminders | *"I just let them know that there's no, um, there are no constraints on what they say."* |
| Implementation of practice sessions | No practice sessions<br>Occasional practice sessions<br>Practice sessions | *"Now I just, um, give them a pen and get them to take it apart of get them to use an unrelated product and just get them to think aloud."* |
| Tasks used during a usability test | Representative tasks<br>Solvable tasks,<br>Task confusion<br>Task design, Task failure<br>Task scenario, Task skipped<br>Verbal tasks | *"So, I would identify what the key, what the key functionality relates to either to the business and to what the business goals are for the product"* |
| Interacting with participants | Minimal interaction<br>No interaction | *"Um, so I'll try and keep my interactions pretty neutral and infrequent"* |
| Interventions during think-aloud session | Immediate prompt<br>Prompt for rich data<br>Stuck prompt<br>Reflective thinking | *"So, if there was, if they made an injection or you know, that there was some element of surprise or confusion"* |
| Participants behavioural change | Exploration<br>Reactivity | *"But I think it might change what they do"* |
| Data Analysis activities | Tasks success rate<br>Session review<br>Metrics<br>Notetaking<br>Results discussions<br>Thematic coding<br>Utterance comparison | *"It will be those tasks that I would focus my analysis activity on. And it would really be at the level of looking back at the videos and what people said and, or, or even checking the understanding that I've taken from the session."* |

### 6.6.1 Think-Aloud Usefulness

#### 6.6.1.1 Expeditious

Practitioners are unaware of the distinct types of think-aloud methods (the classic, retrospective, and Constructive Interaction). The findings from the interview indicate that the concurrent think-aloud method is much more popular than the retrospective and constructive interaction techniques among UX practitioners because they use the concurrent method during usability testing as it helps them to understand participants' tasks-solving strategies, for instance:

*“... And it's just such a quick way of getting at that data and it makes understanding people's experience so much better. You get a much richer picture by asking people to think aloud in the main…” [P1]*

*"...As they're talking through all that, it's helping you understand what's going on in their mind." [P6]*

Also, the vast majority indicated that they had learned TA techniques at work, as well as from UX online/offline bootcamps and use it because of its expeditious nature of accomplishment of getting the tasks done:

*"...Well, yeah, I mean, it makes the, it makes the session so much easier to moderate." [P9]*

*"And a lot of the time given the backlog of work we have; we often don't have that time. Um, so it's more of how do we get the quickest, efficient way to tell the story." [P19]*

*"...based on time and cost and the maximum output." [P8]*

#### 6.6.1.2 Insights

Using think-aloud during usability test help practitioners learn about the users' task-solving strategies this assist them when to learn what goes on inside the head of the users and help them create a product with a better user experience. Many of the interview/practitioners reflected on how the think-aloud method helped them to gain more insights about the tested product:

*"..., I think the think-aloud is the only way that we really just know of finding out what's inside of our participants, Headspace, um, their needs, their motivators, et cetera..." [P8]*

*"...so, we kind of do make sure we get, um, qualitative sort of, um, insights from those situations, but the mere fact that they haven't been able to do it is also an insight for the client." [P20]*

*"But very valuable insight about how people feel about the page about what's important." [P18]*

This further helped practitioner to understand the user's experience as they use the think-aloud during task solving, they were able to understand their challenges and emotions which help them identify the area in the tested product that needs improvement based on the insight obtained from the think-aloud session:

*"And it makes understanding people's experiences so much better. You get a much richer picture by asking people to think aloud in the main." [P1]*

*"And so, when you test and you start taking note of these problems, these are things that you've noticed that you can improve on, on the UI side." [P16]*

*"…which I do find useful that Sometimes more useful just to get thoughts and themes." [P17]*

*"Very interesting information about emotion that's difficult to have with any other metadata." [P13]*

### 6.6.2 How UX Practitioners Use the Think-Aloud Technique During Usability Testing

#### 6.6.2.1 Neutral instructions

Instructions play a major part in usability testing because a test facilitator must clarify what they want the user to do during a usability test. When using classic think-aloud approaches, moderators must only ask participants to express and say aloud whatever comes to mind naturally. Findings from the interview indicate that few facilitators give users neutral instructions:

*"It's me giving them an instruction that I would say is being neutral. So, I would just say, I want you to think aloud. I just want you to say out loud what you would normally say to yourself. I don't ask them to say particular things". [P1]*

*"I would just leave it more. I'll probably just say I'd like you to think aloud as you're going through things." [P15]*

In some cases, practitioners give instructions that are misleading and problematic as such instruction put participants in a reflective thinking mode, for instance:

*"Okay, I want you to search for something that you often search for in this case" [P5]*

#### 6.6.2.2 Building Rapport

Evidence from the study indicates that all twenty-two interviewed UX practitioners build some form of rapport with their users by having a short and friendly conversation with them:

*"So, what I like to do is just have a couple of gentle questions, just see how they're, how they are if I've got any pre-existing information." [P8]*

*"So, like, I always budget like five minutes where we go, how was your day? Where did you travel? Blah, blah, blah." [P7]*

Beyond building a rapport by having a friendly chat, several practitioners also explained that they brief participants about the study to put them at ease and make them feel comfortable to verbalise their thought processes when they are thinking aloud during task solving process within the usability test session:

*"Um, so yeah, I think that's an important part so that people feel comfortable with you in this study." [P1]*

*"Um, I try to diminish the word test in their mind because they, of course, assume they're being tested." [P14]*

*"Yes. I tend to have a brief chat just to make them, uh, more talkative." [P22]*

#### 6.6.2.3 Explicit instructions

Findings from the interview show that practitioners give users explicit instructions:

*"I would just say, um, you know, I'm going to show you something in a minute that we want to get some feedback on and see, you know, see what works for you. Um, it'd be really useful if*

*you could just give me a kind of good and common ratio as you're going through it. Um, we try and keep it a bit kind of, um, a bit relaxed, I guess, a bit flexible". [P10]*

Also, findings from the study show that practitioners are very much in doubt that they might end up with data that is not meaningful to their study:

*"…I could end up with data that's not relevant. And I understand that that does introduce some level of, uh, facilitator bias." [P8]*

Further evidence indicated that most interviewees/practitioners use explicit instructions such as the ones that request participants to verbalise thoughts about their feelings, give reasons for their actions and gives design recommendations:

*"So, uh, what do you feel like when clicking on the button or what are the steps? Uh, so in that way, I tend to, uh, give them instructions." [P4]*

*"Um, so this button is supposed to do this. Do you, is that clear to you thinking out loud, would you click on it all your own? You know, so it depends on what we're measuring intuitiveness design, layout navigation…" [P2]*

#### 6.6.2.4 Reminders

The think-aloud reminder is used to remind the users to "keep-talk" if they fall silent for fifteen to twenty seconds during a usability test:

*"...and I would just when they fell silent, I would remind them to keep going" [P1].*

Contrasting with this, findings from the study indicated that several practitioners use this as an opportunity to prompt and probe participants for explanations for their actions:

*"Um, I think we would just try and not put too much pressure on, but, if there's a long pause, if they were kind of stuck on something, we might say what's happening here or what you think." [P10]*

*" Uh, like nothing specific, I would say, um, just, reminding them. Okay what you've done here, or, um, can you tell me what you're thinking now?" [P12]*

*"Oh, you know, just remember as you're going through, just speak about, you know, what decisions you make, what you are thinking about." [P17]*

### 6.6.3 Implementation of Practice Sessions

#### 6.6.3.1 No Practice Sessions

Part of Ericsson & Simon, (1993) recommendations when using the think-aloud method during usability testing, is to have a short rehearsal for participants to practice think-aloud. However, findings from this study suggested that practitioners don't usually carry out a practice session:

*"…I try not to give them a practice. I just dive into..." [P7]*

*"I wouldn't ask them to practice it" [P11]*

*"No, I don't actually" [P6]*

*"…Uh, no, I probably wouldn't actually ask them to practice." [P15]*

One practitioner explained that they think it not necessary to ask participants to practice a think-aloud session because in all the usability studies they have conducted, their participants do think-aloud with the instructions he gave them:

*" I've just found that it wasn't, hasn't been necessary for me. I've never had a study in which, um, I would say, hey, you know, think out loud as much as you can. Um, and they weren't able to do so if anything," [P2]*

Further findings indicated that some practitioners just verbally explained the think-aloud process to their participants. While another gave a rationale for not asking her participants to practice a think-aloud session as it might make them feel like it's a test that they must pass and not fail:

*"Um, no. I explained what it is." [P14]*

*"So, I feel like I'm, I feel like asking them to practice makes them feel like they have to get it right." [P17]*

Contrasting with this, two practitioners admitted that he has never asked their participant to practice thinking aloud before, however he thinks it could be a good practice to implement:

*"…I have not done that, but I could see it being useful." [P5]*

*“Uh, actually, uh, I don't tend to do that. Um, and that, that could be a good practice, but I don't tend to do that.” [P22]*

#### 6.6.3.2 Occasional Practice Sessions

Findings from the interview reveal that some practitioners occasionally carry out a practice session with their participants if they felt that the participant might experience some difficulty thinking aloud during a usability testing session then they will conduct a practice session:

*"Like I couldn't do that, but I would only do that if I was talking to someone who had really minimal context on the idea" [5]*

*“...Um, have you done it once or twice, especially if I think the use that might need that little bit of extra guidance." [P3]*

*"A few times, but not really. I think, if you explain it well enough, then, I think that people really do understand. I mean, I've done it a few times, but really, I find that just explaining, what we mean by think-aloud, um, just does the jobs, so we don't really need to do a demonstration and practice of it yet...." [P10]*

One practitioner explained that he tends to omit the think-aloud practice session due to the time factors: *"It depends. I'll tell you what it depends on time. So, um, if I were to look at percentages, um, and stay in the business world, think-aloud, I would always try and put it in there, but practically realistically, it would probably be about 40% of the time." [8]*

#### 6.6.3.3 Practice Sessions

Findings from the interview session indicated that some practitioners still conduct a practice session as it prepares the participants for the task ahead and to ensure that they understand what it might be to think aloud by giving them a demonstration of how to think aloud:

*"I would give them a demonstration and maybe ask them to do it, for example. Yeah." [P10]*

*"Yes, I do. I just, I used to get them to play a mine sweeper and think aloud on that. But then I, for some reason I stopped doing that. Now I just, um, give them a pen and get them to take it apart of get them to use an unrelated product and just get them to think aloud. And I make sure that I explain to them quite carefully what I want them to do." [P1]*

*"I always have a test session, created a usability think aloud just to get the participant to, to practice early on." [P8]*

### 6.6.4 Tasks Used During Usability Test

#### 6.6.4.1 Representative Tasks

Findings indicated that most practitioners make use of representative tasks. Usability testing is a task-based approach, thus the type of task used during a usability test play a significant role and influences the outcome of the test. Evidence from the study shows that practitioners use representative tasks that users are expected to perform on the digital product or to find specific information:

*"...if it's usually like, e-commerce, it'd be like, how do you, um, make a purchase?" [P2]*

*"Just representative tasks for, for the product that we're working with." [P1]*

*"...So, what you have to do is like perform the key task of the digital product." [P4]*

*"…so, in those examples, I was giving the one about the scheduling tool. It was like, I want you to make a schedule for your team." [P5]*

*"So, it's a very specific, specific thing, uh, usually to see if they can find this specific thing or perform some action with this specific item."[P12]*

#### 6.6.4.2 Solvable Tasks

Findings from the interview shows that practitioners usually give participants solvable tasks, these are tasks around the tested product aims and functionalities:

*"...Yeah. And I wouldn't give them anything that they couldn't answer with the product. So, I would identify what the key, what the key functionality relates to either to the business and to what the business goals are for the product." [P1]*

*"So, I get, I try the main objective is to get the participant to think and imagine that they're in that situation." [P8]*

*"...I try and let them, you know, do something that's real to them." [P17]*

*"...what we want to have at the end of, of our usability test session. And then I asked them to try to implement that." [P22]*

#### 6.6.4.3 Task confusion

Even though practitioners give their participants representative and solvable tasks. Practitioners also explained that something participants struggles to complete a task due to misunderstanding about interface tasks and then requesting for help from the test facilitator because they are unsure if they are doing the right thing:

*"...And you can see they're scrolling all the way down and back up again very quickly, but why are they scrolling down an app? Is it because they are still not sure that is the right thing they shouldn't be doing. " [P6].*

*"I'll try and ascertain. If the reason they're struggling is perhaps the scenario isn't clear enough." [P3].*

*"Sometimes the body language doesn't correspond with what they're actually saying. like, where's the shirt Where can I find this shirt?" [P7].*

Task confusion is a problematic aspect of usability test because when participants are unsure of what to do, they stop think-aloud, which sometimes leads the test facilitator to prompt them to explain their difficulty: *"I see that sort of look of concentration and then I see them stop talking because they're confused." [P14].*

#### 6.6.4.4 Task Design and Task Scenario

Finding from the interview reveal how practitioners design the task that they use for usability testing. Practitioners explained that they design task in scenario so it will be fun and fairly real for participant to accomplish:

*"So, it could be a hotel website. And we'll say, um, imagine you're taking, um, a relative on holiday for the, you know, it's a surprise gift, go and select a room for two weeks." [P20].*

*"...um, the basically it's a kind of scenario going through a few steps and whatever it is that we're testing, I suppose. Yeah." [P15].*

Practitioners also reflect on creating both open and close-ended tasks that are broken into smaller bits that are less complex and achievable:

*"And um, I think we always try to make sure that the task is not Was its cumbersome that doesn't even, it's like this giant thing." [P119].*

*"...uh, I make more, uh, more tasks and two types of tasks also close-ended and open-ended." [P12].*

Findings also indicates that when most practitioners design a task, they tend to be describing the tasks to make it simple and comprehensible for participants:

*"So, imagine your, uh, you want to go out hiking and you need to prepare a gear for that, and go ahead and, uh, and put some things in the basket that you will find useful on such trip." [P12].*

*"So, I will give them a scenario, which is normally you've come through. You've claimed something within two to three weeks..." [P3].*

*"… it was like, we made up this story that you're the manager and your supervisor have to approve the schedule." [P4].*

*"...Um, I make sure that it's short as possible, but it's got enough context, uh, make sure that it's scenario based in the story of, uh, um, uh, in the form of a story or a scenario." [P8].*

#### 6.6.4.5 Task Failure

Practitioners explained comments where participants struggle to accomplish a task due to poorly designed product and moment where participants failed to complete a task and yet announces that they have completed the task:

*"So even if four out of five were able to find it, the three of them can be found it because trust me, users will see the executed or the tasks that you could clearly see. They did not complete." [P7].*

*"... it was, I don't know, it was a bit heart-wrenching watching these people because they were struggling so bad, and these are chartered accountants." [P14].*

These findings indicate how UX practitioners use the think-aloud to gain a deeper insight about task failure rate and its implications which are used to inform design recommendations.

#### 6.6.4.6 Task Skipped

Findings from the interview reveal some situation when the test facilitator asked participant to skip a task:

*"...Or sometimes if, worst case scenario, I can just skip that task and ask participant go to the next one." [P19].*

#### 6.6.4.7 Verbal Tasks

Findings from the interview indicated that some practitioners give participants verbal tasks. Although, one might argue that verbal task makes a usability test more conversational and having a usability test in a conversational manner makes a usability test seems like an interview. However, some practitioners prefer giving verbal tasks to their participants:

*"I prefer to actually give verbal commands and actually say, all right. Okay, good. And the whole point is that you need to make them feel that it is a progression." [P6].*

*"So, across those five tests, we got those as a baseline. Um, the person moderating the task will basically just read them out. We never give them on a piece of paper." [P20].*

One practitioner reported reading the tasks to participant, while another indicates that she sometimes writes the tasks down for participants but often she usually read it out to them:

*"So usually, I sort of read it out, read it out to them." [P19]*

*"Um, but there might've been an instance where I had things written down, but, um, yeah, usually the verbal." [P11]*

Some practitioners reported that, he usually read the tasks verbally then hand the written task to the participants for reference purpose:

*"I'd probably read it and then give it to them. And so, they can refer back to it." [P17].*

*"So, uh, unless there's something, would I need it to be the same for every person. I will just read it out if it wants it to be exactly that would maybe have it written for them." [P15]*

### 6.6.5 Interacting with Participants

#### 6.6.5.1 Minimal interaction

When using the think-aloud during usability testing, it is often suggested limiting interaction with participant to only think-aloud reminder “please keep talking” and that interaction with participant should be kept to a minimal level. Findings from the study revealed that although most practitioners reported they tend to limit their interaction with participants. However, they do prompt their participants to give explanations to their actions at some points:

*" I try not to probe too much during the think aloud." [P17].*

*"Say that I prompt them though. I pretty well will leave them to their own devices" [P21].*

*"Um, so I'll try and keep my interactions pretty neutral and infrequent because the test isn't about me, it's about their interaction with the, with the system " [P1].*

*"...Um, I try to minimize some of my interaction" [P3].*

*"I'm instructing someone to think aloud, I want to avoid talking for more time than if it was in natural normal conversation" [P5].*

*"... Uh, you know, generally I just wait and listen to them, you know, listen to them." [P6]*

In contrast to the above comments, one practitioner reported that, he does not interact with participants because it might alter their task-solving strategies:

*"Um, I will try and keep them going without really interfering too deeply because I know that can, can be problematic." [P1]*

### 6.6.6 Interventions During Think-Aloud Session

#### 6.6.6.1 Immediate prompt

Findings from the study indicated that, practitioners use immediate prompts. This implies that participants are prompt in an immediate manner, instantly or without delay:

*"I noticed that you said this. Can you tell me more about that, that kind of thing to try to keep it in the frame of mind of what they're saying, their language, as opposed to my language or my company's language.”? [P5].*

*"In case they're navigating to some different flow altogether, then you have to like, uh, as participant, why did they do that?" [P4].*

Practitioners reported the reason for their immediate prompt:

*"...And then, because if I waited until the end, there might forget what it was, or I might take it into the wrong place." [P1].*

*Also, practitioners also reported that, sometimes participants carry on with their task solving without thinking aloud or verbalising their thought:*

*"For example, a participant is doing something and clicking and is not saying why. And you're like, okay, stop right now." [P13].*

*"Maybe you can tell from body language, if they kind of go like a bit surprised or something, you just say, Oh, what were you thinking?" [P15].*

*"Um, but also if I am the moderator, I tend to ask questions on the go, uh, on the fly." [P12].*

*"Um, perhaps literally, um, uh, or what are you thinking right now?" [P11].*

#### 6.6.6.2 Prompt for rich data

Findings from the interview indicated that most think-aloud reminder used my practitioners during usability test is a prompt for rich data. When participant is stuck, instead of using the simple think-aloud reminder "keep talking", practitioners prompt and further probe participants for explanation and to give design recommendations:

*"So, you clicked on this, but you didn't tell me that you were about to click, and you didn't tell me why." [P13].*

*"So, they would click here, click there. Sometimes we might pause them and say, oh, I saw you, you were trying to click this, or you click this, you know, what exactly did you see that made you do that?" [P19].*

*"So, if I feel like there's something useful, that's come up that it would be useful to talk about in the moment then. Yeah. I'll, probe that." [P17].*

*" I would probe with questions, you know." [P11].*

*"We would, once they've, you know, we'd say what happened there, what you think, and they would answer, and then we would basically respond with something." [P10].*

*"You know, if they do a certain thing, we'll ask them to explain it in words" [P20].*

#### 6.6.6.3 Stuck prompt

During usability test, one of the instances where test facilitator should prompt their participants is if they are stuck or when there is a task confusion. One practitioner reported that they only interfere with participants task solving process if they felt that the participant is stuck, and he/she is not moving forward with the study:

*"Um, I think when, when you need to, when I would, mainly interfere would be, if somebody was stuck on a task" [P1].*

Practitioners also reported that, when a participant is stuck, they sometimes use it as a medium to obtain “rich data” from participants by further probing them to give verbal explanations to the reason why they are stuck:

*"If worst case scenario is getting stuck, like it's a, and he's not able to go forward, he's not able to understand. Then you can intervene and ask, okay, this is hmm" [P4].*

*"I would leave them to kind of do it as much, but then I would maybe prompt that point where I think, well, they're stuck here or something like that" [P15].*

*"You know, once they finish the task or if they get stuck or if they have any questions, then that's when we kind of come in" [P10].*

Findings from the study reveal that while most practitioners will probe their participants for explanations, other tries to reassure the participant when they are stuck and encourage them to carry on with the test:

*"And then that's what I'll just step in to say you know, do you feel like you're totally stuck? I mean, I've got that from understanding by prompting them to think a lot more, um, that they are totally stuck, then I will unstuck them. [P14]*

*"Say, you know what, don't worry about it. I mean, as I told you, we are just testing, is it, this is not complete. Are you okay about trying to do something else?" [P6]*

*"So honestly, tell me if you're stuck or if I wouldn't use the word, I would say something like, um, are you able to proceed?" [P7]*

#### 6.6.6.4 Reflective thinking

Findings from the interview indicated that practitioners do prompt participants for explanation to their actions. Such prompt alters participants thought process and put them in a reflective thinking mode, where they must access their long-term memory instead of the verbalising their thought as it occurs to them. For instance, see some comments from some of the interviewed practitioners:

*"If I, if I feel like something's going on and they didn't phrase it. Uh, so I asked them." [P22].*

*"Why did you have that reaction? And it just, you know, feeds into a conversation." [P21].*

*"Um, for them to describe how they feel about what they're seeing and to describe what they expect to happen." [P2]*

One practitioner reported that, asking participants questions during task performance might change their task solving strategy:

*"Well, um, if you start asking people questions that might get them to, to change their strategy. Um, so I think inadvertently you could nudge them into a different task solving strategy, which could then either, um, lead to better performance or it could derail them." [P1].*

Similarly, another practitioner reported that he probes participant, knowing that probing for rich data or information from a participant during usability test could be problematic as it alters their task solving process:

*"I do probe, and I understand that that does introduce some level of, uh, facilitator bias." [P8]*

*"Um, can you explain what you're thinking right now?" [P7]*

### 6.6.7 Participants Behavioural Change

#### 6.6.7.1 Exploration

The interview explores if the act of think-aloud makes users want to explore more. Here are some of the comments from UX practitioners:

*"Hmm. Yes. I think that some participants, uh, um, might be bias to explore more because they want to convey a lot of information because of the think aloud process." [P22].*

*"I think they can be more like slightly more exploratory and willing to explore when thinking out loud." [P12].*

*"Um, yeah, probably. Yeah. It' not a kind of reliable method in that sense that it's not naturally what people would do, but it's a good kind of indication" [P10].*

*" So, I think it makes people try harder." [P7].*

One practitioner reflects on his experience that, the fact that users sometimes get paid to explore a tested product, often makes them wants to explore more and sometimes give design recommendations:

*"...definitely the whole setup offer, like the research setup, the fact that you're testing participant in the lab or testing them in general, uh, often to get paid for that, it's increasing their willingness to explore and to work on tasks they would abandon otherwise." [P12].*

*So, um, I don't know. I think maybe the thinking aloud might actually prompt them to kind of poke around a little bit." [P14].*

Two practitioners reported that in addition to the fact that the think-aloud get users a little bit hyped where they try to accomplish the tasks give to them, the artificiality of a usability test setup is a contributor to make them explore more than they would normally do:

*"Yeah. I think that, but I think that's, that's, um, a risk of research as a whole. I don't think it's this technique alone." [P17].*

*"The thing aloud will make them explore a bit more. Also, just the environment of doing testing, the sort of artificiality of doing a usability test or research will make them explore more." [P3].*

#### 6.6.7.2 Reactivity

The interview explores practitioners view on reactivity: a change in task performance which makes a usability test no longer a representation of real-world use. Findings indicated that the think-aloud do put users in a state of artificiality and sometimes make them wants to explore more. Here are some of the comments from practitioners:

*"But I think it might change what they do, because if they're having to keep that's that stream going and giving that inclination, they might, they might take a broader view of the page." [P1].*

*"There is a chance that it can put somebody in a state, more artificial mindset and forced them to be looking for more things or try more stuff that they wouldn't normally do." [P3].*

*"I expect it probably does. To some extent there is a sort of artificial sort of nature of the setup." [P11].*

One practitioner reported that, sometimes test facilitators unknowingly pressurise users because they want to get feedback or some sort of validation from the user:

*“I think that we, as researchers might push people for like, yes or no answers to quickly, because we might be pressured to like validate something or work really fast." [P5].*

In addition to the above comment, three practitioners also reported that the cognitive workload of having to carry out task performance and think-aloud simultaneously is problematic especially during difficult task as people may try to think-aloud over their thoughts and this can have an impact on their behaviour:

*"We ask people to actually talk over their thoughts. So that is a cognitive issue. There that's actually makes the thinking aloud, uh, already a little bit problematic." [P6].*

*"And even when they get stuck, that you try and voice out what is going through their mind." [P7].*

*"...but yes, I would see that the think-aloud does absolutely impact on their, you know, behaviours." [P8].*

Three practitioners reported that, sometimes user worried about time factor which makes them get very anxious and that has an impact on their behaviour and how they navigate the tested product while think-aloud, however were unsure of the extend of such impact on their behaviour:

*"So, I guess, there's a risk that it might affect what people naturally do by a small percentage maybe, but not so much that it's gonna, you know, uh, negate your findings." [P10].*

*"Not drastically. I don't think, I think because you've given them a task to do, they would still be following that in a sense." [P15].*

*"Obviously that has an influence on how they, I guess, navigate or how you think or the impression. Um, so I would definitely say there is, to what extent do is what I don't know." [P19].*

One practitioner is of the opinion that, the think-aloud sometimes do make users overthink during tasks performance:

*"So, there's a chance they could overthink it" [P20].*

### 6.6.8 Data Analysis Activities

#### 6.6.8.1 Notetaking and Data Analysis

Findings from the study reveal how UX practitioners analyse the data they obtained from a usability test. While some practitioners indicated that they no longer carry out notetaking activities because they are recording the test session. For instance, see comment from one practitioner:

*"...I stopped taking notes. No, because I'm recording the session or I'll watch it back, you know, at least once, probably twice." [P9].*

Contrary to the above comment, two practitioners reported that, they use notetaking techniques to obtain valuable insight from participants during usability testing. Here are comment from practitioners:

*" Sometimes people will in their thinking aloud will actually try to ask questions out loud and I will tell them I can't answer anything. Um, but I'm making notes" [P14].*

*" I take note of it, and I say, hey, this user couldn't recognize this feature straight away" [P16].*

Others indicated that they still take note of instances where they observe something very interesting, they are able to write them on the spot and dig deeper into that later on and instances where users encounter difficulties for ease of access when they go back and replay the recordings:

*"...if it's just a level of easy completed and some difficulty, and then where are the areas of difficulty then you just go back to the video or very quickly, sometimes I'll note time and I'll just quickly go back" [P1].*

*" So, I reviewed all video is again, uh, and the first reactions, instantaneous reactions, uh, how people startles, struggled" [P12].*

*"Even if I I'm doing it a week later, I will then watch the video. And that's when I start capturing, okay, this was an act, because I can pause the video. I can look at exactly what screen they were on. I can catch all my data points" [P7].*

*"I might write something that I do not have the bucket for, I'll take note of that because it's so unique to that user." [P7].*

*"Um, it's really useful to be there because it cuts down the time because I can make a note there. And then on each task, I usually have a notebook and the tasks written out for myself and I'll jot down notes." [P1].*

One practitioner reported using transcribing their data and using coding as a way of analysing the frequencies of users' verbalisations:

*"Um, if I'm doing it for my research, uh, I've got a set of codes that have been extracted from the literature, and I will be looking to say, um, what the relevant frequency of those codes were within the transcript." [P1].*

Two practitioners reported grouping data into main themes to look for cause and consistent trends to gain useful insights from the data they obtained from a usability test:

*"And what I would do is I would do it on a granular level. So, I, um, I would group into main themes, and I would have a set amount." [P8].*

*"And then I start finding cause and consistent trends or what we call thematic insights." [P7].*

*"Um, generally get them to all take five or 10 minutes just to put on post-it, you know, the key things that stood out to them and then we'll have a conversation and theme it and then talk about like what potential actions are." [P17].*

One practitioner reported using task success to inform their recommendations:

*"… there were 10 participants for the study. So, seven out of 10 participants thought this is what this means. And three participants never understood because so, and so reason. And so, screen level analysis we do, and screening level design recommendations are also done" [P4].*

#### 6.6.8.2 Utterance Comparison

Only two practitioners reported that they carry out utterance comparison from different user to get a clear picture of the data they obtained from a usability testing which help them to gain a deeper understanding of what users are saying. Here are some of the comments from practitioners:

*"So, we would separate observations from what people are saying." [P15]*

*"Uh, then the whole discussion about how many t people said that, how many people said this and that" [P18].*

*"we'll compare and, you know, make sure we've got the full version of anything if we need to..." [P20].*

#### 6.6.8.3 Results Discussions

Findings from the study reveal how UX practitioners create their report from the data they obtained from a usability test. Practitioners reported that more often they get their team members involved and discussed the results they obtained using discussion guide, the goads and objective of the study and client requirements. Here are some of the comments from practitioners:

*"And so, we can basically summarize then the clients get the choice of having just a summary version of that report, or they can have a more detailed report" [P20].*

*"...I'll always have a discussion guide for my session, so that's got, key questions in there and then there's always going to be key goals or objectives for the research." [P17].*

*"... And then after all the testing, when we all get about together with it again, back to the scale. And that's really discussion and debate kind of thing." [P14]*

*"Um, traditionally I've always tried to get my team involved." [P3]*

## 6.7 DISCUSSION

The present study explored the practices and challenges of using the think-aloud protocol in the industry. The study aims to answer the following research questions: (i) How do UX practitioners use the think-aloud method within usability testing? (ii) What is the nature of tasks practitioners' uses? (iii) What are practitioners' views on reactivity?

The results of the study provide insights into the practices and challenges of using the think-aloud protocol in the industry, as well as the underlying rationale for the think-aloud administration procedures. Findings from the study confirms some of the concerns identified in the literature, nevertheless, there are others where the results contradict prior reports.

To answer the underlying research questions, this discussion focuses on five main themes: (i) the think-aloud usage, particularly the prevalence of how it is been administered (ii) tasks (iii) instructions for requesting verbalisation (iv) intervention this include prompting and probing and (v) how practitioners analysed the data they obtain. In the subsequent paragraphs the author will discuss results and their implication in detail.

### 6.7.1 Think-Aloud Usefulness

Findings from the study concerning the concurrent think-aloud usage indicated that the concurrent think-aloud method are widely used in both remote and controlled lab usability studies because it gives insights into people’s expectations. These findings were in accordance with previous studies conducted by McDonald et al., (2012); Fan et al., (2020).

Indeed, some of the qualitative comments from practitioners suggested that it makes the session so much easier to moderate as a quick way to get insights and it makes understanding people's experience so much better. However, the method of usage of the think-aloud technique varies widely among UX practitioners.

Previous research has shown that there is a gap between theory and practice when it comes to using think-aloud methods (Boren and Ramey, 2000; N0rgaard and Hornbeek, 2006; Shi, 2008; McDonald et al., 2012). A study conducted by Fan et al., (2020) shows that this gap remains. Similarly, findings from this study are consistent with Boren & Ramey, (2000); and Fan et al., (2020).

The subsequent section will focus on four important themes. (i) rapport building (ii) think-aloud instructions (iii) practise sessions (iv) task design and task types used in think-aloud sessions and (v) how practitioners analysed the data they obtain.

### 6.7.2 How UX Practitioners Use the Think-Aloud Technique During Usability Testing

#### 6.7.2.1 Rapport Building

Findings indicated that all 22 practitioners establish some sort of rapport with their participant before a usability test. Although not consistent with how it is been done, however, the objective was always to put participant as ease. Some practitioners give participants assurance beyond what is reasonable by giving specific details about the test and what to expect. Also, after building rapport, most practitioners continue to reinforce to the participants that they are not been tested, that is the product or interface that is been tested.

According to Sonderegger and Sauer (2009) a good rapport between the moderator and the test participant may also improve performance. Hence, there should be a way for practitioners to stick a balance between establishing a good rapport and giving specific details about a test or going beyond what is reasonable in order not to influence their behaviour.

#### 6.7.2.2 Think-aloud Instructions

When using concurrent think-aloud protocol within usability testing, test facilitators must only ask participants to express and say aloud whatever comes to mind naturally (Ericsson and Simon, 1984). Boren and Ramey (2000) found that test facilitators did not follow Ericsson and Simon's guidelines when giving instructions to participants.

Findings from the present study, indicated that only two of interviewees (practitioners) adhere to Ericsson and Simon's recommended guidelines, while most of the interviewed practitioners adopt a more specific or explicit instructions like those used by their superior in their workplace, or instructions they learned from their superior and recommendations from UX Research bootcamp or usability textbooks (Barum, 2010, Rubin & Chisnell 2008; Dumas and Loring 2008).

Evidence from the study further show that, the rationale behind practitioners' use of explicit instructions outside Ericsson and Simon's recommended guidelines was fear that they could end up with irrelevant data and they are perceived to provide rich data as opposed to the procedural explanations that are often depicted in research studies. This is consistent with previous research such as (Cook, 2010, Zhao and McDonald 2010 and Boren and Ramey (2000).

Also, Findings from this study indicated that only one out of the twenty-two interviewed practitioners reminded their participants to keep talking when they fall silent for 15 to 20 seconds without actively probing them with questions while they were thinking aloud. Similarly, a study conducted by Fan et. al (2020) shows that only 16% of their respondents reminded their participants to keep talking without prompting and further probing them with questions during think-aloud. Ericsson and Simon's think-aloud guidance has been replaced by explicit instruction in certain usability testing textbooks. Barnum (2001), for example, suggested a

think-aloud exercise in which participants express their mixed experiences, explain their choices, and offer reasons for their utterances.

Further evidence from this study indicates that practitioners are very much in doubt that they might end up with data that is not meaningful to their study. For example, one participant said: “Hopefully there will be something in there. Um, if I find that second the content or the way that they're speaking is not valuable in terms of when I go to code or analysis or thematic analysis later, then I find that, um, I gently kind of persuade them in that area. And I understand that does introduce some level of facilitator bias.” [P12] and “I could end up with data that's not relevant,” [P18].

This finding is consistent with previous studies which stated that some of the reported instructions and interventions were not required (Nerqaard and Hornbeek, 2006) and unskilled interventions could add bias and lead to erroneous data and practitioners are aware of these concerns (McDonald et al., 2012). This indicates that some practitioners are aware of the level of bias regarding the type of instructions they give to participants but are more dreadful of not obtaining desired data so they are opt-in for instructions that will lead them to obtain rich data and intervene if needed.

#### 6.7.2.3 Implementation of Practice Sessions

Findings indicated that most of the practitioners that were interviewed (86%, n=19) indicated that they do not ask their participants to practice think-aloud, 14% (n=3) only do it occasionally and that depends if they sense that the participant might struggle with thinking aloud, and only 9% (n=2) out of the 14% (n=3) do it almost all the time. Moreover, only 9% (n=2) gave an example of the practice session they used.

Evidence from the study indicated that, the rationale for not conducting a think-aloud practice session were because most practitioners find that just explaining what they meant by think-aloud to participants is effective. So, they don’t see the need for a practice session. While

some practitioners see time as a big overriding factor, others use the practice session when participants are confused about thinking-aloud.

This result is consistent with a study by McDonald et al. (2012) which found that usability practitioners consider think-aloud to be a simple concept for users to understand. In their study, 52% of respondents indicated they never use a practice session, and 15% only conduct a warm-up on rare occasions. Indeed, previous study within usability testing Fan et al., (2020) reported that only 24% of their respondents use a practice session when conducting a usability test. These findings are consistent with this current study.

#### 6.7.2.4 Tasks Used During Usability Test

Findings from this study show that, practitioners gave users or participants solvable tasks mostly in the form of task scenarios which depend on the product that is been tested. Findings from the study also showed that practitioners are more concerned about task completion and spend a significant amount of time on task design especially when working with team member they come together and anticipants where they think users might encounter difficulties when using the product.

Findings also indicates that the way practitioners administered task to participants varies widely and can be categorise into three types of approaches: (i) those that give verbal task as a form of scenario, (ii) some hand in the tasks to participant for them to read and (iii) others read each task to participants and then hand it to them to read or refer to during a usability test session.

This implies that, most practitioners know that task is an imperative aspect of a usability test, however, are less concern about its administration procedure which might accidentally influencing the participant's behaviour. However, this is not evident in previous research within the field of usability testing and could be an interesting area for further research. Although, in

terms of task type and reactivity, previous research has found no link between task-types and reactivity (second study on this thesis).

#### 6.7.2.5 Interacting with Participants

Findings from this study indicated that in terms of interactions, practitioners indicated that they kept their interaction with users to a minimal level. However, this does not indicate that they used to recommend guidelines to only ask participants to “keep talking” when they fail silent for 15 – 20 seconds without further probes as suggested by Ericsson and Simon. Rather they probe when participants fall silent and further prompt for rich data.

For example: "So they would click here, click there. Sometimes we might pause them and say, oh, I saw you, you were trying to click this, or you click this, you know, what exactly did you see that made you do that?" [P19]. This finding is in accordance with a previous study that reported that test facilitators use interventions other than think-aloud reminders (Fan et al., 2020; McDonald et al., 2012; Nørgaard and K. Hornbæk, 2006; Shi, 2008 and Boren & Ramey, 2000). Similarly, “if participant is stuck like he is not able to go forward, then I can intervene and so I can further probe him as per where do you think you should click.” [P20]. On the Contrary, “If there was something that happened that if they asked me a question, I would answer it in such a way as not to steer them. Um, I will try and keep them going without really interfering too deeply because I know that can, can be problematic.” [P1].

Ericsson and Simon (1993) emphasised the importance of interacting with participants as little as possible. To reduce thought process disruptions, the experimenter should only issue think-aloud reminders if participants go silent, and the reminders should be brief and non-directive, such as "keep talking." Interactions between the experimenter and the participants would place the latter in a social context. This may encourage participants to use level 3

verbalisations and formulate their comments using higher-order reasoning to ensure they are understood, increasing the likelihood of reactivity. Evidence shows that this is not the case in the industry.

#### 6.7.2.6 Interventions During Think-Aloud Session

Concerning facilitators Intervention and types, findings from the study indicated that practitioners disclose that their intents are utilised to get a deeper understanding of participants' utterances, behaviours, or responses to the interface, which is sometimes required by the context of the test. Findings also, showed that, practitioners intervened and interact with participant like it's a cognitive behavioural therapy (CBT). For example: ", I prompt them, but it's more like a reflective practitioner technique used in CBT counselling, which is okay. I see you're struggling there. Can you tell me a little bit more about what's happening? I'm also aware of my words and how I influenced the participant and I want to keep the depth" [P17].

These findings are consistent with previous study conducted by McDonald et al. (2012) which find that that when it came to interfering, most usability professionals took a modified strategy. McDonald et al. (2012) found that interventions were determined by the participants' characteristics, as well as the scenarios that arose during the testing, and, on occasion, the client's requirements.

In contrast, to the above comment from participant 17. Three out of twenty-two participants don't think that intervention during think-aloud changes users' behaviour. For instance: "Um I don't think it necessarily changes their behaviour. Um, and not from my experience. I don't believe that is necessarily altered how they have gone through this. But they're still going to click over there and click on. I found that doesn't change what they click on."

However, previous studies indicates that, evaluators' interventions may involve urging users to go beyond their current situation and consider hypothetical scenarios (Nerqaard and

Hornbeek, 2006; Shi, 2008). They may also include indications that point users in the right direction for completing a task (Nerqaard and Hornbeek, 2006).

Similarly, “I won't say it changes people's behaviour; I don't think it changes their behaviour in their ability to find the things to execute the task.” [P7]. However, previous studies indicated that interventions may interrupt users and change their behaviour at the interface. For example, study conducted by Hertzum et al. (2009) measured test participants' mental effort, eye movements, interface behaviours, and task-based performance. They compared the impacts of classic think-aloud and relaxed thinking aloud to working in silence. Participants took longer to finish tasks and engaged in more extensive online browsing, scrolling activity because of the relaxed approach.

They speculated that these effects were due to the participants' increased awareness and insecurity about their task-solving strategies or that the evaluator's interventions made them less able to concentrate. Participants performed better when evaluators intervened, according to Olmsted-Hawala et al. (2010a). They wanted to see if employing the traditional think-aloud, speech-communication technique versus the coaching approach (relaxed think-aloud) had any effect on participants' standard task performance, specifically time on task, task accuracy, and satisfaction with the site. As a control variable, participants who worked in silence were used.

The researchers discovered that proactive interventions enhanced task solving accuracy and resulted in participants having a substantially greater level of satisfaction with the website than in other situations. Concerning Prompting participants, findings from the study indicated that practitioners prompt participants during think-aloud session and their prompt can be categorised into three main types of prompts: immediate prompt; stuck prompt and prompt for rich data.

Findings from the study shows the rationale behind practitioners' immediate prompt. For instance, "And then, because if I waited until the end, they might forget what it was, or I might take it into the wrong place." [P1]. Similarly, "We prompt them at that point and because you've got that person in that situation, like doing that thing for that one time only." [P22].

In a previous study, conducted by Petrie and Precious (2010) participants were asked to convey their views about a website while undertaking a variety of tasks. Findings showed that these inquiries produce information about users' emotional responses. Such prompts are an indication of reflective thinking which is contrary to Ericsson & Simon established guidelines.

According to a previous study by Nielsen, (1993) UX practitioners may deviate from Ericsson & Simon recommended guidelines and engage with their participants in two instances. One is when participants are caught in a difficult situation. In this case, interacting with them to assist them in recovering from the issue would allow the test to progress as such allowing UX practitioners to find other usability issues. Another scenario is when participants are dealing with a well-known issue, the impact of which has been discovered and well understood by prior test participants. It's less useful to sit and watch people struggle with the problem over again. However, practitioners go beyond these two instances in the quest for rich data as evident in this study.

Concerning the stuck prompt, findings from the study indicated that practitioner's observer and ensure participants are stuck entirely and then intervene to un-stuck them. For instance, "Um I think when you need to when I would mainly interfere would be if somebody was stuck on a task" [P1]. Similarly, "…are you able to proceed to execute the task?" [P7] and "I don't say anything unless he hit a wall where they're stuck entirely, I will un-stuck them" [P4]. These comments from participants 1, 4 and 7 are in accordance with previous research conducted

by Nielsen, (1993) which states that UX practitioners may deviate from Ericsson and Simon (1993) guidelines and engage with their participants when participants are caught in an inconvenient situation.

Concerning When they prompt for rich data, findings from the study indicated that, practitioners prompt their participants when they fall silent, and further probe them for explanation regarding their actions during task execution. For instance: “so you clicked on this buy you did not tell me that you were about to click, and you did not tell me why. So, they explain why.” [P13]. Similarly, “I might try to sort of nudge them and prompt them a little bit to share what’s going through their mind at that time.” [P11].

Additionally, “If they go silent for 20 to 30 seconds, then I ask them you know what they’re thinking, what they’re trying to do.” [P6]. According to McDonald et al. (2012), 71% of usability practitioners adopt a flexible approach or always intervene while obtaining thinking aloud data. Also, among the group with the most experienced respondents, the proportion of those using a flexible approach was the largest.

Despite Ericsson's and Simon's recommendations that practitioners utilise neutral instructions, Fan et al., (2020) found that only 7% of respondents followed this recommendation. Most responders specifically requested that their participants express other types of data, such as feelings, emotions, activities, and even design suggestions.

A rationale behind this divergent practice could be associated with previous studies. For instance, Makri, Blandford, and Cox (2010; 2011) have stated that classic think-aloud is insufficient for delving deeply into users' behaviour since users frequently fail to give reasons for their actions. Similarly, Cooke (2010) discovered that the traditional technique mostly provided procedural descriptions (such as reading or describing activities), which are less useful for usability assessment (Boren and Ramey, 2000).

This is problematic since studies that urge participants to verbalise a specific type of content might alter their task-solving behaviour, thus masking usability issues (McDonald & Petrie, 2013).

#### 6.7.2.7 Participants Behavioural Change

Findings from this study reveals that the type of instruction that is been used by a test moderator might change participants behaviour by putting them in a state of more artificial mindset and force them to be looking for more things or try more things that they would not normally do. Findings also shows that, practitioners opine that, they are a risk that it might affect what people naturally do by a small percentage and unsure of extent of such implications.

This finding is consistent with Ericsson and Simon's suggestion, that generic instruction should be substituted for think-aloud instructions. Ericsson and Simon (1993, 16 p.80) suggested that the former often included "complementary" bits for extra information, such as requests for explanations or the elicitation of certain categories of content, such as instructions to provide likes and dislikes. Ericsson and Simon warned that these might alter the structure of participants' mental processes by drawing their attention to specific information and forcing them to engage in the self-monitoring process.

#### 6.7.2.8 Data Analysis Activities

Findings from this study indicate that UX practitioners turn to their observation notes and specific areas of the session recordings more often than playing back the entire video of the recorded session or transcribe user' verbalisation. A rationale behind this is based on time factor. Previous study within the field of usability suggested that UX practitioners often work in agile environment and face time pressure and budget constraint (Rose and Cardinal, 2018).

Indeed, evidence from this study is in line with these findings. Although, most practitioners are aware that using only their observation note and not referring to the session recording might sometimes lead to inaccurate result, however, they often must make trade-offs between meeting deadlines and achieving high reliability and validity. Also, UX practitioners has developed better note taking techniques where their colleague takes note while they moderate the think-aloud session, and then do a team discussion of the session and produce a report. However, it unclear whether this method of data analysis is robust or if other method is available.

## 6.8 Limitations

In order to mitigate against bias possibly introduced by interviewing and coding combined by the author was reduced by 3 weeks break between running the entire interview session and the thematic data analysis procedure. Although, it would have been better to use separate researcher for each activity. Also, the study was limited to UX practitioners in the UK. Future research could investigate such practice and challenges in other locations other than the United Kingdom.

## 6.9 Summary

The focus of this research was to fill a gap in our knowledge of UX practitioners' practises and challenges when employing the think-aloud approach in the industry by combining theoretical viewpoints found in the literature with the dynamic influences of UX practitioners' current experience in the UX industry. An interview session with 22 UX practitioners in the United Kingdom formed the basis of the research. The research uncovered the practises and challenges that come with using and analysing think-aloud sessions. The findings from this study indicated that, the concurrent think-aloud method is widely used in both remote and controlled lab usability studies, UX practitioners establish some sort of rapport with their participant before a usability test and most UX practitioners do not conduct a think-aloud practice session as they consider time as an overriding factor while other only do a practice session when they participant are confused about thinking-aloud.

Findings also indicates that the way practitioners administered task to participants varies widely and can be categorise into three types of approaches (see details in discussion section). Findings shows that, UX practitioners use explicit instructions during usability testing and that, practitioners probe participants when they fall silent and further prompt them to explain their actions and to verbalise specific type of content. Evidence from the study further shows that, practitioners prompt especially when a participant is stuck or struggling, to unstick the participant they prompt for rich data.

Evidence from this study reveals that the type of instruction that is been used by UX practitioners might change participants behaviour by putting them in a state of more artificial mindset and force them to be looking for more things or try more things that they would not normally do. Concerning behavioural change, findings from the study shows that, practitioners opine that, they are a risk that the way the think-aloud has been used in addition to test

facilitators intervention during task performance, it might affect what people naturally do by a small percentage but was unsure of the extent of such implications.

The findings of this study are significant to both education and practise, and they have the potential to enlighten UX practitioners and academia about how their peers view and use the think-aloud protocol, as well as inform academia about practitioners' use of the think-aloud protocol. The findings also point to ways to improve some of the strategies for conducting and facilitating a usability test when using the think-aloud technique.

# CHAPTER SEVEN

## 7.0 ANALYSIS AND DISCUSSION

### 7.1 Overview

This thesis has examined the causes of reactivity of the concurrent think-aloud: the first study provided evidence to support previous research that the classic think-aloud is not reactive. The second study looked at differences in task performance and the third study explores practitioners' use of concurrent think-aloud, their practises and challenges.

This chapter outlines the major insights presented in this thesis, and summarises the answers to the research questions, discusses the research limitations and provides recommendations for future research.

### 7.2 The Reactivity of the Concurrent Think-aloud within Usability

The goal of research on the use of think-aloud techniques during usability testing has been to identify the elicitation circumstances that lead to reactivity — an unnatural, frequently favourable shift in task performance caused by the concurrent think-aloud (Hertzum, et al., 2009; McDonald and Petrie, 2013 and McDonald, Zhao and Edwards 2015).

These elicitation techniques involve giving specific instructions or using interactive prompts to elicit verbalisations that activate higher-order cognitive functions like elaboration and explanation (Ericsson and Simon, 1993; Fox, Ericsson, & Best 2011). A reactive think-aloud is problematic during usability testing since it jeopardises the reliability of the collected data. Evidence from study 2, see chapter five for details.

### 7.2.1 The Classic Think-aloud and Reactivity

The idea of reactivity and its connection to the work provided in this thesis should be discussed before answering the research question. Numerous researchers have been examining the reactivity effect linked to the concurrent think-aloud, as was mentioned in section 2.4. According to study conducted by Fox et al (2011), if Ericsson and Simon's recommendations are carefully adhered to, the concurrent think-aloud will not induce reactivity. In other words, the think-aloud technique will not cause a change in participants tasks performance either improves or diminished. On the contrary, deviating from Ericsson and Simon's framework by requesting verbalisations through targeted instructions or interactive prompts that activate higher-order cognitive processes like elaboration and explanation may result in reactivity.

The first research question in this thesis is:

**RQ1** (i): Does act of thinking-aloud under classic administration procedures causes reactivity?

The Concurrent think-aloud has been known to cause reactivity. Another name of this phenomenon is the self-explanation effect (Chi et al, 1994). It's an artificial change, (enhanced or diminished performance) in task performance making the test no longer representative of real-world use Fox, Ericsson, and Best (2011).

This is problematic because it may alter the accuracy of task performance, thus leading to poor usability problem detection, low data reliability and validity. Ericsson and Simon, (1980, p. 27) stated that, "the accuracy of verbal reports depends on the procedures used to elicit them" and reactivity will occur when the established procedure is neglected. Empirical demonstrations of reactivity within usability testing have shown mixed findings Olmsted-Hawala, et., al (2010) and studies show that practitioners don't follow these guidelines (McDonald, Zhao, and Edwards 2015).

The results from the first study on this thesis strengthen Ericsson and Simon's (1993) theoretical framework by demonstrating the 'non-reactivity' of the concurrent think-aloud, as the result from the study shows that, there was a significant main effect of task ($F(1,18) = 38.576$, $p<0.001$). Overall, participants spend more time when completing Sensemaking tasks (mean= 329.97) than when completing fact tasks (mean= 150.58). There were no other significant main effects or interaction. The result accords with (Ericsson and Simon, 1980; Van den Haak, de Jong and Schellens, 2009; Ericsson and Fox, 2011; McDonald and Petrie, 2013; McDonald, Zhao and Edwards, 2015) and discords with (Wright and Converse, 1992; Van Den Haak, De Jong and Jan Schellens, 2003; Eger et al., 2010) who's studies found the CTA to be associated with reactivity.

The results reaffirm that, the act of thinking aloud under classic administration procedures does not cause reactivity of the concurrent think-aloud techniques within usability testing.

The next section presents the impact of different task-types: fact and sensemaking tasks on the concurrent think-aloud techniques and its influence on reactivity.

### 7.2.2 The Impact of Different Task-types to Influence Reactivity Within Usability Testing

Tasks plays an important role in usability testing and different task-types are used by test facilitators which are often classified as simple and complex tasks. As detailed in section 2.6, task-types includes fact finding, assessment and sensemaking tasks.

To answer the research question RQ1 (ii) this study makes use of fact and sensemaking tasks.

The second research question in this thesis is:

**RQ1 (ii):** Does the use of different task-types: fact tasks and sensemaking tasks have impact on the reactivity of the concurrent think-aloud?

According to Lewis (2001a), even the habit of assigning only achievable tasks in usability testing may bias the findings. Hornbk (2006) advocates for tasks that demand cognitive problem-solving rather than purely motoric behaviours. Users would have a greater chance of evaluating system's utility as well as its user-friendliness this way. Therefore, not informing participants when a task is complete could lead to them incorrectly believing the task is completed when it is not, according to Boren and Ramey (2000). This has a practical implication in the form of usability test facilitator intervention, which could disrupt the participant's usual task flow.

To begin with, requiring users to think aloud while performing a task can result in a dual processing effect in which cognitive resources are split between the task and the think-aloud protocol. Secondly, task performance suffers due to the necessity to think aloud (Chin and Schooler 2008).

Evidence from this study demonstrates that the act of thinking aloud under classic administration procedures does not cause reactivity within usability testing and task type does not have impact on the reactivity of the concurrent think-aloud, although, sensemaking task lead to an increase in mental demand and effort. There were significant differences for the following subscales (independent t-test): Mental Demand (t= .315, df = 18, p=.041); Performance (t= 1.529, df = 18, p=.046); Effort (t=.641, df = 18, p=<0.014). see details in table 4.8.

Findings from study one implies that task-type does not have impact on the reactivity of the concurrent think-aloud as the study has no evidence which indicate a change in participants' behaviour. Although, the type of task in which this effect is most likely to manifest are cognitive demanding tasks such as sensemaking[2] tasks as evident from the study (chapter four) shows

[22] sensemaking tasks involves effort to understand relations in order to select a best course of action.

that sensemaking task led to an increase in mental demand and effort as a result of increased concentration and attentional demands. See details in table 4.8.

### 7.3 Effect of Explicit Instructions on Participant Task Performance

The third research question in this thesis is:

**RQ2(i)**: What is the impact of task performance on the use of fact and assessment task with the classic think-aloud, explicit instruction or silent within usability testing?

According to this study, using explicit instructions with task types such as fact and assessment tasks is more problematic as the classic think-aloud provide useful indication in terms of its impact on task performance. Nonetheless, explicit instruction resulted in more verbal utterances in categories that usability practitioners expected and found relevant. On the other hand, participants who received explicit instruction, completed fewer tasks successfully, and the explicit instruction resulted in a higher mental workload in terms of performance than the classic think-aloud technique. Nain effect of task ($F(1,57) = 1288.372$, $p<0.001$), participants spend more time during ETA than on CTA when completing Assessment tasks (mean=193.80) than when completing fact tasks (mean=128). For number of clicks ($F(1,57) = 585.126$, $p<0.001$); number of additional pages ($F(1,57) = 382.867$, $p<0.001$); number of correct tasks ($F(1,57) = 598.162$, $p<0.001$), participants completed more fact tasks (mean= 0.65) than Assessment tasks (mean= 0.49). For number of abandon tasks ($F(1,57) = 141.858$, $p<0.001$); number of partly completed tasks ($F(1,57) = 46.658$, $p<0.001$) and number of incorrect solutions ($F(1,57) = 88.316$, $p<0.001$). see table 5.4 for details.

These findings are in accordance with study conducted by McDonald and Petrie (2013), explicit instructions for users to indicate their likes, dislikes, and confusions cause people to

navigate and scroll more often than when working silently. As evidence from the study indicated that, when completing the assessment task, participants in the explicit condition reported higher levels of frustration than those in the classic and silent conditions.

Overall, findings from this study indicates, Data elicited from participants through the use of explicit instruction may be a false representation of the user's interaction with the tested product. Furthermore, previous research has suggested that it may be a contributing factor to likely cause reactivity (Gerjets et al., 2011; Fox at al., 2011).

### 7.3.1 The Impact of Explicit Instruction

The fourth research question in this thesis is:

**RQ2(ii)**: Does explicit instruction lead to high mental workload over classic think-aloud and Silent?

Most of the research conducted in usability test domains back up the Erisson and Simon (1980) model. In the investigations by Hertzum et al. (2009), for instance, the test users' performance is slower in the conditions with thinking aloud compared to silent working, but with classic thinking aloud, no additional effect is seen. On the other hand, when compared to working silently, relaxed thinking aloud alters users' behaviour in various ways by boosting general web page surfing and reported mental effort. Additionally, the findings of Wright and Converse (1992) reveal notable variations between the test groups.

The results showed that explicit instruction resulted in a higher mental workload in terms of performance than classic think-aloud and silent conditions, but classic think-aloud and silent resulted in a higher mental workload in terms of effort. (Performance: CTA, 71.75 (15.79); ETA, 74.65 (14.30); and Silent 63.35 (18.00)), see table 5.5 for details. Although effort was more noticeable in the silent condition during the assessment task, this suggests that participants may have exerted extra effort because they were performing task performance in silence.

In terms of the explicit instructions, findings indicated that participants did not have to put in extra effort due to the explicit instructions that they received. This finding is in accordance with finding from McDonald et al. (2013a), which indicated that explicit instructions may aid in the completion of difficult tasks. Hence, participants didn't put in more effort.

### 7.3.2 The Role of Explicit Instructions within Usability Testing Regarding Explanatory Utterances

The fifth research question in this thesis is:

**RQ2(iii)**: Does explicit instruction lead to an increase in relevant explanatory utterances in terms of user experience and expectations?

In studies on thinking aloud from the 1960s, participants are asked to explain each decision they make when solving the Tower of Hanoi problem or other comparable tasks, rather than merely verbalising their thoughts in their working memory (Gagné & Smith 1962; Davis et al. 1968). According to the frameworks of Ericsson and Simon (1980), thinking aloud in these studies is therefore on level 3, where it is claimed that doing so might alter participants' normal behaviour.

Also, these studies demonstrate that participants who had been thinking aloud during the preceding tasks do better on the final, more difficult tasks.

According to the Ericsson and Simon (1980, 1993) model, thinking aloud has no influence on performance at level 1 or level 2, with the possible exception of slowing things down. In these levels, participants verbalise their ideas from working memory without giving reasons for their decisions or actions. For instance, this hypothesis is supported by research on problem solving by Rhenius and Deffner (1990), which found that thinking aloud did not affect problem-solving accuracy but did increase tasks completion time.

In this study 20 participant carried out task performance using the classic think-aloud techniques; another 20 participants were given explicit instructions and the other 20 performed their tasks in silent. Less time is spent on tasks by participants in the thinking-aloud condition, and they also make fewer errors. Time on tasks: Silent condition, Fact tasks, 139.80 (38.39), Assessment tasks, 194.40 (34.73); CTA, Fact: 123 (40.06), Assessment 192.60 (48.85); and ETA Fact, 121.20 (35.58), Assessment, 194.40 (53.46). See table 5.4 for details.

This impact is also evident when the tasks are more difficult, such as an assessment test. Thus, the model developed by Ericsson and Simon is once more supported by the changes in user behaviour.

Overall, the classic think-aloud condition produced more verbal utterances than the explicit condition. Although, the explicit condition produces more utterance category for “user experience” and “expectations” compared to the classic condition. This finding is in accordance with study conducted by McDonald, McGarry and Willis, (2013) Which indicated that there were proportionally more explanatory utterances as a result of the explicit think-aloud. Similarly, Zhao et al., (2012) which stated that, the participants in the explicit instruction condition reported a higher mental workload and a concentration on identifying interface issues, but there were no differences in task performance. Additionally, the explicit instruction condition produced more user experience, expectation, and behaviour explanation statements than the neutral condition.

Indeed, evidence from the study also showed that, despite not being directed to, participants in the classic condition ended up verbally evaluating the site's user experience and making recommendations.

The next section presents the UX practitioners’ practices and challenge when using the think-aloud within usability testing.

## 7.4 Practitioners' Use of Concurrent Think-Aloud

The sixth research question in this thesis is:

### 7.4.1 Practitioners Use of the Concurrent Think-aloud Method Within UX Industry

**RQ3** – What are the practices and challenges of using the think-aloud protocol in the industry?

This study examines practitioners' experiences, views on reactivity and challenges when using the think-aloud method within usability testing to make an informed decision when using the think-aloud protocol within usability test and to explore methods to address these challenges. This was further broken down into three research questions which are outline below:

### 7.4.2 Implications of Modifying the Traditional Concurrent Think Aloud Methodology

The seventh research question in this thesis is:

RQ3(i) How do UX practitioners use the think-aloud method within usability testing?

Evidence from this study qualitative analysis reveal that practitioners understand the importance of building rapport with their participants as the results indicated that all practitioners establish some sort of rapport with their participants. In terms of think-aloud instructions, evidence from the study reveal that practitioners often use a more specific or explicit instruction when using the think-aloud techniques, instead of asking participants to express and say out loud what comes to mind naturally as recommended my Ericsson and Simon, (1984, 1993).

The study also reveals that most of the practitioners do not ask their participants to practice think-aloud. This finding is in line with a study by McDonald et al. (2012) that discovered usability experts believe users can easily understand the notion of think-aloud. Also, only 24% of respondents, according to Fan et al. (2020), use a practise session before a usability test. These findings coincide with the current research.

Evidence also indicated that practitioners utilise the explicit instruction for fear that they could end up with irrelevant data and that the explicit instructions are perceived to provide rich data as opposed to participants verbalisation of procedural explanations. This finding is consistent with previous research such as (Cook, 2010, Zhao and McDonald 2010 and Boren and Ramey (2000).

It's crucial to interact with participants as little as possible, according to Ericsson and Simon (1993). The test facilitator should only remind participants to think aloud if they stop talking, and the reminders should be brief and non-directive, such as "keep talking," in order to minimise thought process disruptions. Participants would be placed in a social context through interactions with the test facilitator. This would encourage participants to utilise level 3 verbalisations and higher-order reasoning to guarantee that their comments are understood, which would increase the likelihood of reactivity. Evidence suggests that this is not the case in the UX industry, as practitioners reported that they prompt and further probe for explanations and their intents are utilised to get a deeper understanding of participants behaviour. Practitioners also reported that they do sometimes interact with participants like it is a cognitive behavioural therapy. However, previous research suggests that evaluators' interventions can involve persuading users to think beyond their present circumstance and take hypothetical scenarios into consideration (Nerqaard and Hornbeek, 2006; Shi, 2008). They could also provide clues that direct users to specific locations for carrying out tasks (Nerqaard and Hornbeek, 2006).

Findings also reveal that practitioners prompt participants when they fall silent and further probe them for explanation instead of reminding them to “keep talking” as recommended by Ericsson and Simon. This finding concurred to study conducted by Fan et. Al (2020) which shows that only 16% of their respondent reported reminding their participants to keep talking without prompting and further probing them for explanation to their actions. The study also reveals that practitioners prompt can be categorised into three main types: immediate; stuck and prompt for rich data. Participants in a study by Petrie and Precious (2010) were asked to

provide their opinions about a website while completing a range of tasks. Findings demonstrated that these inquiries reveal data about users' emotional reactions. Such a challenge suggests reflective thinking, which is contrary to the accepted standards set by Ericsson & Simon.

Previous research could be linked to a rationale for this divergent practise. According to Makri, Blandford, and Cox (2010; 2011), classic think-aloud is insufficient for delving deeply into user behaviour because users frequently fail to provide reasons for their actions. Cooke (2010) discovered that the classic think-aloud primarily provided procedural descriptions such as reading or describing activities, which are less useful for usability evaluation (Boren and Ramey, 2000).

In terms of using the think-aloud during usability test, practitioners reported two major challenges: one is the challenge of creating a neutral environment that invites people to honestly express their thought processes. However, most practitioners lessen this by building rapport at the beginning of a test session. Keeping participant interaction to a minimum is another challenge that practitioners have identified as some practitioners reported that such intervention could introduce some level of bias to the study and keeping the interaction to a minimal level comes with experience. This finding is consistent with study conducted by Fan et al., (2020) which listed three significant challenges respondents had when conducting a usability test using the think-aloud technique.

### 7.4.2 The Role of Task/Task-type in Usability Testing

The eigth research question in this thesis is:

**RQ3(ii)** What is the nature of tasks practitioners uses?

The importance of task to a usability test cannot be over emphasised as detailed detail in section 2.6, usability test is a task-based process. The results of this study demonstrate that practitioners primarily presented users or participants solvable tasks in the form of task scenarios which further depends on the tested product. The study's findings also revealed that practitioners spend a lot of time on task design and are more concerned with task completion. One important finding from this study is the way practitioners administered task to participants which varies widely among practitioners. Some give verbal task as a form of scenario, others hand in the tasks to participant for them to read or refer to during a usability test session.

### 7.4.3 UX Practitioners' Views on Reactivity of the Think-aloud protocol

**RQ3(iii)** What are practitioners' views on reactivity?

In terms of practitioners view on reactivity, the results of this study show that the sort of instruction provided by a test facilitator may alter participants' behaviour by forcing them to adopt an artificial mentality and force them to look for more things or try more things than they would typically do. Additionally, the study also reveal that practitioners are unaware of the full scope of such consequences but believe there is a small chance they could somewhat alter what people naturally do.

This result supports the idea made by Ericsson and Simon that generic instructions should take the place of think-aloud instructions.

According to Ericsson and Simon (1993, 16 p. 80), general or explicit instructions frequently contain "complementary" requests for more information, such as clarification requests or the elicitation of specific types of content, like instructions to offer likes and dislikes. By focusing participants' attention on certain details and compelling them to use the self-monitoring

process, Ericsson and Simon cautioned that these could change the participants' mental processes.

## 7.5 Original Contribution

The work presented here has contributed to pushing the boundaries of knowledge in the application of the concurrent think-aloud within usability testing both in research and in practice.

### 7.5.1 Knowledge Advancement In The Application Of The Concurrent Think-Aloud Techniques & Reactivity

***The reactivity of the concurrent think-aloud***

This research is the first to systematically investigate the impact of task-types on reactivity of the concurrent think-aloud within usability testing.

This research makes the unique contribution of encapsulating the value of classic concurrent think-aloud and its variations with usability evaluation using various task-types, such as fact and sensemaking tasks and a detailed verbal analysis using categories that were specific to participants' user experience and usability.

The work builds upon valuable literature in usability testing as findings from empirical studies shows mixed findings of the reactivity of the concurrent think-aloud techniques. Having to combine think-aloud with task performance causes reactivity (Van den Haak, et al., 2004). However, Ericsson and Simon, (1980) identified elicitation procedures that were not associated with Reactivity and stated that reactivity will occur when the established procedure is neglected. The study in this thesis supports Ericsson and Simon’s research and shows that the classic think-aloud is not Reactivity.

This research used categories that were unique to the setting of usability evaluation to define the sorts of utterances made at a comprehensive level, in contrast to earlier studies which examined the verbal data at a general level. This research provided information about the usefulness of verbal data obtained from a usability study, this makes it possible for the researchers and practitioners to more precisely describe the variations in the kinds of

utterances produced by various technique applications (as shown by the findings) and to highlight their importance to usability practitioners' objectives.

#### 7.5.1.1 Recommendations

- Researchers and practitioners should make every effort to carry out a practice session prior to the actual usability study, as this will allow participants to practice and become accustomed to verbalising their thoughts more frequently, reducing the need for test facilitator intervention or prompt for action explanations.
- The use of neutral instruction, as suggested by Ericsson and Simon (1985, 1993), should be properly adhered to Instead of seeking rich data or asking participants to provide a certain sort of context.
- Practitioners should keep interaction with participants to a minimum and should only intervene for task clarification or if a participant becomes stuck on a task.
- When participants are quiet for 15 – 20 seconds without verbalising their thoughts during task performance, instead of probing and prompting for more detailed information, practitioners should utilise the recommended think-aloud reminder as suggested by Ericsson and Simon, to encourage them to "keep talking" without any further prompt or additional instructions.

### 7.5.2 Insight into the Trades-Off Involved in Eliciting Level 3 Verbalisation and Reactivity

This study makes a significant contribution to bridging the gap between the use of the classic think-aloud and explicit instruction. For instance, studies within usability testing have documented divergent practice in the use of think-aloud instructions and test facilitators interventions and have compared both classic and relaxed think aloud or explicit instructions, for instance, Hertzum et al., (2015), investigated verbalisation in usability test by comparing participants verbalisation in moderated and un-moderated test during relaxed think-aloud. A study conducted by McDonald and Petrie, (2013) investigated whether the

classic think-aloud and a think-aloud with an explicit instruction led to different task solving performance compared to silent working. Similarly, Zhao et al., (2014) compared the classic think-aloud and an explicit instruction requesting explanations and content that is relevant to the user experience. McDonald et al., (2013a) investigated whether an explicit explanation-based think-aloud instruction leads to differences in navigation performance over the classic think-aloud method. Cooke, (2010) study addresses the use of think-aloud protocols in usability test settings with respect to users' verbalisation accuracy, verbalised content, and what do users' eye movements reveal about their behaviour when they are silent.

Also, Zhao and McDonald, (2010) compared the classic and a relaxed think-aloud with the aim to explore the impact of think-aloud style on the nature of the utterances produced by participants and the usefulness of those utterances for usability analysis. However, none has compared Classic think-aloud, the use of Explicit instruction and Silent working with Fact, and assessment tasks to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity.

While earlier studies (e.g., Hertzum et al., 2009; Olmsted-Hawala et al., 2010) demonstrate that departing from Ericsson and Simon's theoretical framework, particularly the use of explicit instructions and intervention can affect the validity of data, the practical implications of these deviations have not been investigated. However, this divergence practice is the most important factor that influences technique a usability practitioners chooses. The divergent perspectives of academics and practitioners are a result of their respective focuses. Academics eagerly caution practitioners to the consequences of changing an approach without strong theoretical foundation. On the contrary, these techniques have been continuously employed by practitioners, and in fact the gap between research and practise is expanding rapidly (Fan et al., 2020; McDonald et al., 2012).

Given the current situation, this study makes a significant contribution to bridging this gap for usability practitioners to understand the trade-offs involved in eliciting level 3 verbalisation and test reactivity. The results informed usability practitioners that, although, the explicit instruction did produce more verbal utterances in categories that usability practitioners expected and find relevant, the explicit instruction participants completed fewer tasks successfully and led to high mental workload in terms of performance. The results informed usability practitioners that the supposed benefits of the use of an explicit instruction do not seem to outweigh the risk.

### 7.5.3 Practitioners' Use of Concurrent Think-Aloud: Practises and Challenges

The results of this study have implications for both education and practise. They may help UX practitioners and academia better understand how practitioners utilise the think-aloud protocol and how peers perceive it. The research presented in this thesis has shown that the use of the concurrent think-aloud techniques varies widely among practitioners and the method used to elicit the think-aloud can have an influence on what participants do while carrying task performance.

Findings also indicates that the way practitioners administered task to participants varies widely and can be categorise into three types of approaches: (i) those that give verbal task as a form of scenario, (ii) some hand in the tasks to participant for them to read and (iii) others read each task to participants and then hand it to them to read or refer to during a usability test session.

Ericsson and Simon (1993) emphasised the importance of interacting with participants as little as possible. The findings demonstrate that UX practitioners utilise explicit instructions during usability testing, and that practitioners probe participants when they go silent, prompting them to clarify their actions and verbalise certain kinds of information. Practitioners reveal that they face the fear of not getting valuable data from participant. Hence, they have to prompt for action explanation and probe to get desirable verbalisation.

The results also reveals that the type of instruction that is been used by UX practitioners might change participants behaviour by putting them in a state of more artificial mindset and force them to be looking for more things or try more things that they would not normally do. Practitioners are faced with the challenges of not prompting participant and ending up with irrelevant data.

### 7.5.4 UX Practitioners' Views on Reactivity of the Think-aloud protocol

This research is the first to systematically explores UX practitioners' views on reactivity. In terms of reactivity, findings from the study reveals that, practitioners opine that, they are a risk that the way the think-aloud has been used in addition to test facilitators intervention during task performance, it might affect what people naturally do by a small percentage but was unsure of the extent of such implications. This is an important direction for future research.

#### 7.5.4.1 Recommendations:

- Practitioners should design their research to capture and identify patterns emerging from participant data (e.g., verbalisations, actions, emotions, eye-tracking) that typically occur when users encounter usability difficulties. This will assist to reduce the need to interfere, prompt, or interrogate participants.
- Practitioners should recognise the tradeoff between obtaining high validity data and resolving low severity usability issues from a think-aloud session. As a result, more effort should be put into research design, particularly task design and recruiting representative participants, in order to attain high data validity and reliability.
- Practitioners should improve their notetaking skills to get the most out of a usability session. When observing participants during a think-aloud session, practitioners should take notes of timelines or timestamps when participants encounter difficulties since these instances provide vital insights to the research and will help in improving

the efficiency and communication of usability analyses and reduce the need for intervention.

- Practitioners should pay attention to participants' actions and verbalisations, such as pitch and speaking pace, when analysing think-aloud sessions. This will assist them in identifying usability concerns.

### 7.5.5 The Research contribution to UX practice and Research

This study makes a significant contribution by demonstrating that the link between research and practise is more complex than just arguing against adopting research, which has been the main topic of discussion among academics up until now. The study's results, however, showed that practitioners employ explicit instruction in a more opportunistic way.

Another contribution challenges the perception that practitioners are not interested in theory or methodologies for usability testing that have been established via research. However, practitioners stated that these approaches must be in resonance with their own experience of usability testing practise and their own perceptions of how professionals utilise the method. In contrast, usability test procedures with the concurrent think-aloud, in the opinion of practitioners, are not sufficiently grounded in the reality of practise. Therefore, the issue with academically created usability testing with the use of the concurrent think-aloud protocol is not necessarily that they are too theoretical or too abstract, but rather that they fail to take into account the realities of everyday usability testing practise.

Findings suggest that practitioners are interested in and seek research-based usability testing procedures.

This cycle emphasises the importance of studying about the usability testing process and how practitioners use concurrent think-aloud in a more direct way to fully understand when and how traditional procedures are used, as well as how the procedure is modified based on the context of the test and the practitioner's judgement.

## 7.6 Limitations

This PhD research, like any research of its kind, has some drawbacks that could be addressed by future research. This section discusses the general study limitations that apply to all three investigations. The pertinent chapters address the study limitations that are particular to each study.

The research could have explored other causes of reactivity by examining the impact of test facilitator's presence on think-aloud to influence reactivity. However, the goal of this research is to provide practitioners with results that are helpful for their job as well as usability research. In order to avoid focusing on an overlooked procedural component, the author chose to concentrate on the methodological changes made by the practitioners and their use of concurrent think-aloud: practises and challenges. This does not imply, nevertheless, that the practise is not significant, and more study in this area is needed.

Secondly, the coding of the qualitative data by the author. Within the framework of a PhD thesis, this was unavoidable. However, steps were taken to limit any possible bias. Coding was cross-checked to ensure participant verbalisations and verbal evidence for studies one and two without being aware of the conditions from which the data was derived. While a lecturer from a different department who is well-versed in qualitative analysis and not been involved in the study other than the coding reviewed the coding data for study three. With respect to study one and two, a reliability check of breakdowns was also in place, but due to the size of the data set and the second coder's restricted availability, it was not possible to completely analyse all the data.

Another limitation is that the author could introduce bias when conducting the usability test. According to Clemmensen et al. (2012, 2009), the evaluator's cultural background is likely to influence the results of the usability testing in areas like think-aloud instructions and tasks, the evaluator's reading of the user, and the interaction between the user and evaluator as a whole.

There was a chance that the participants' behaviour and think-aloud data would be impacted because the author is of a different nationality than the participants. But the author has long resided in the UK and pursued his studies here.

Finally, the research didn't consider how the methods' application at different stages of the design/development lifecycle could influence approach and data. For instance, whether the methods were applied during exploratory design, interactive prototyping, acceptance testing, or competitor reviews could bias the results. Future research is required to address these limitations.

## 7.7 Future Work

This research identifies various future research directions.

First, the study detailed in section 4.2 investigated the impact of task type in terms of complexity on the reactivity of the CTA within the context of usability testing. The study uses two different task-types: fact and sensemaking tasks. Future research should investigate the value of level 3 verbalisation.

Secondly, the study detail in section 5.2 builds upon previous studies within the field of usability testing to investigate the impact of task-type on two different think-aloud protocols and its effect on participant task performance, test experience and verbalisation by comparing the classic think-aloud, explicit instruction and silent working with fact and assessment tasks.

Concerning the implications of explicit instructions, more information is required. In particular (i) the nature of the instructions such as: if the instruction request is only to think-aloud, or to provide specific information or explanations and recommendations, (ii) the timing of the request to think-aloud, i.e., whether participants are informed prior to task execution at the start of the study or whether they are informed at the start of each task. Also, Future research could look at the trad off in using explicit instruction during usability test.

Finally, this study detailed in section 6.2 examines practitioners' experiences, views on reactivity and challenges when using the think-aloud method within usability testing. However, the study only focused on UK based UX practitioners. Future research could investigate such practice and challenges in other locations other than the United Kingdom.

## 7.8 Conclusions

The most popular usability testing technique may be thinking-aloud protocols, although descriptions of this technique in the usability literature and the working practises of practitioners do not match the theory that is most frequently referenced for it: Protocol

Analysis: Verbal Reports as Data, a landmark publication by Ericsson and Simon (Ericsson and Simon, 1984).

Evidence from usability literature indicate that the concurrent think-aloud procedures vary widely among usability practitioners (Fan et al., 2022; McDonald et al., 2012). Having to combine thinking aloud with task performance causes Reactivity (Van den Haak, et al., 2004). Ericsson and Simon, (1980) identified elicitation procedures that were not associated with Reactivity. They formulated a guideline for the concurrent think-aloud protocol and stated that reactivity will occur when the established procedure is neglected. Several other usability evaluation publications use the term "thinking aloud" in connection to Ericsson and Simon (1993), but without reliably differentiating between verbalisation at levels 1 and 2 and verbalisation at level 3 (p.23).

Also, investigations of what usability evaluators do in practise indicate that relaxed thinking aloud is widespread, and that the rich data acquired in this manner are valued by usability practitioners (Boren and Ramey 2000, Nrgaard and Hornbaek 2006). The reactivity of the concurrent think-aloud is problematic because: It may alter the accuracy of task performance ; Poor usability problem detection; and Low data reliability and validity.

Findings from empirical studies shows mixed findings concerning reactivity (p.81). Therefore, this research has used a different exploration approach to examine the concurrent think-aloud protocol and reactivity by investigating the impact of task-type on the influence of the concurrent think-aloud to cause reactivity. This allowed the research to address the underlying research questions while also providing insights into a possible future for the concurrent think-aloud.

This research has conducted three different studies to investigate the reactivity of the think-aloud and to understand the practices and challenges with the conduct and analysis of think-aloud sessions.

This research clearly shows that we should not forgo concurrent think aloud since the methods created valuable think-aloud data for usability study and helped identifying usability issues. Also, we should not neglect Ericsson and Simon's think-aloud framework in favour of an explicit instruction. As findings from this research reveals that, using explicit instruction had far more substantial effects on participants' behaviour than traditional think-aloud. As a result, Ericsson and Simon's classic concurrent think-aloud method should be used to collect concurrent data not solely because it aims to ensure the validity of the data collected, but also because it generated the same type of data when explicit instruction is used, both in terms of explanatory and user experience data.

This research has summarised the results of an interview study with a range of UX practitioners on the practice and challenges of using the concurrent think-aloud method within usability testing. The research identified several key areas where method use departs from traditional framing in the HCI literature, including an increased in the use of explicit instruction. This research provides several important directions forward which inform future research on the use of the concurrent think-aloud method and the issue of reactivity, with implications for the research and practice to support usability testing practice.

The author suggests that the research's findings will expand our understanding of how to employ the concurrent think-aloud protocol and foster a dynamic interaction between appropriation, abduction, and situated action that will open up a number of new research directions. Finally, the finding of this research will aid UX practitioners in understanding how their coworkers view and apply think-aloud protocols.

## REFERENCES


Abdel Latif, M. M. (2018) 'Using think-aloud protocols and interviews in investigating writers' composing processes: combining concurrent and retrospective data', International Journal of Research & Method in Education, pp. 1–13.

Ahmad, N., Shoaib, U. and Prinetto, P. (2015) 'Usability of Online Assistance From Semiliterate Users' Perspective', International Journal of Human–Computer Interaction, 31(1), pp. 55–64. doi: 10.1080/10447318.2014.925772.

Albayrak, D. and Cagiltay, K. (2013) 'Analyzing Turkish e-government websites by eye tracking', in Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2013 Joint Conference of the 23rd International Workshop on. IEEE, pp. 225–230. Available at: http://ieeexplore.ieee.org/abstract/document/6693243/ (Accessed: 13 February 2017).

Albert, W. and Tullis, T. (2013) Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes. Available at: https://books.google.co.uk/books?hl=en&lr=&id=bPhLeMBLEkAC&oi=fnd&pg=PP1&dq=Usability+measurement+in+context&ots=R8RihuUZtF&sig=VVdOMO1bWpbb_N6OyrzM3n0Ge04 (Accessed: 8 February 2017).

Alhadreti, O. (2016) Thinking about thinking aloud: an investigation of think-aloud methods in usability testing. University of East Anglia. Available at: https://ueaeprints.uea.ac.uk/id/eprint/61487 (Accessed: 13 February 2017).

Alhadreti, O. et al. (2014) 'The Impact of Usability of Online Library Catalogues on the User Performance', 2014 International Conference on Information Science & Applications (ICISA), p. 1.

Als, B. S., Jensen, J. J. and Skov, M. B. (2005a) 'Comparison of Think-aloud and Constructive Interaction in Usability Testing with Children', in Proceedings of the 2005 Conference on Interaction Design and Children. New York, NY, USA: ACM (IDC '05), pp. 9–16. doi: 10.1145/1109540.1109542.

Als, B. S., Jensen, J. J. and Skov, M. B. (2005b) 'Exploring verbalization and collaboration of constructive interaction with children', in IFIP Conference on Human-Computer Interaction. Springer, pp. 443–456. Available at: http://link.springer.com/chapter/10.1007/11555261_37 (Accessed: 8 February 2017).

Alshamari, M. and Mayhew, P. (2009) 'Technical Review: Current Issues of Usability Testing', IETE Technical Review, 26(6), pp. 402–406. doi: 10.4103/0256-4602.57825.

Alshammari, T., Alhadreti, O. and Mayhew, P. (2015) 'When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods', International Journal of Human Computer Interaction, 6(3), pp. 48–64.

Altuntaç, P. and others (2015) The Comparison of Concurrent and Retrospective Think Aloud Methods in Unmoderated Remote Usability Testing. Available at: https://openaccess.leidenuniv.nl/handle/1887/34669 (Accessed: 13 February 2017).

Aranyi, G., van Schaik, P. and Barker, P. (2012) 'Using Think-aloud and Psychometrics to Explore Users' Experience with a News Web Site', Interact. Comput., 24(2), pp. 69–77. doi: 10.1016/j.intcom.2012.01.001.

Arsal, G., Eccles, D. W. and Ericsson, K. A. (2016) 'Cognitive mediation of putting: Use of a think-aloud measure and implications for studies of golf-putting in the laboratory', Psychology of sport and exercise, 27, pp. 18–27.

Bainbridge, L. (1979) 'Verbal reports as evidence of the process operator's knowledge', International Journal of Man-Machine Studies, 11(4), pp. 411–436.

Ball, L. J. et al. (2006) 'Applying the PEEP method in usability testing', Interfaces, 67(Summer), pp. 15–19.

Barendregt, W. et al. (2006) 'Identifying usability and fun problems in a computer game during first use and after some practice', International Journal of Human-Computer Studies, 64(9), pp. 830–846. doi: 10.1016/j.ijhcs.2006.03.004.

Barendregt, W. et al. (2006) 'Identifying usability and fun problems in a computer game during first use and after some practice', International Journal of Human-Computer Studies, 64(9), pp. 830–846.

Benyon, D. (1993) 'Accommodating individual differences through an adaptive user interface', Human Factors in Information Technology, 10, pp. 149–149.

Bergstrom, J. C. R., Olmsted-Hawala, E. L. and Jans, M. E. (2013) 'Age-Related Differences in Eye Tracking and Usability Performance: Website Usability for Older Adults', International Journal of Human–Computer Interaction, 29(8), pp. 541–548. doi: 10.1080/10447318.2012.728493.

BEVAN, N. (1995) Software Quality Journal, 4(null), p. 115.

BEVAN, N. and MACLEOD, M. (1994) 'Usability measurement in context', Behaviour & Information Technology, 13(1–2), pp. 132–145. doi: 10.1080/01449299408914592.

Bezerra, C. et al. (2014) 'Challenges for Usability Testing in Ubiquitous Systems', in Proceedings of the 26th Conference on L'Interaction Homme-Machine. New York, NY, USA: ACM (IHM '14), pp. 183–188. doi: 10.1145/2670444.2670468.

Bias, R. G., Moon, B. M. and Hoffman, R. R. (2015) 'Concept Mapping Usability Evaluation: An Exploratory Study of a New Usability Inspection Method', International Journal of Human–Computer Interaction, 31(9), pp. 571–583. doi: 10.1080/10447318.2015.1065692.

Boren, M. T. and Ramey, J. (2000) 'Thinking Aloud: Reconciling Theory and Practice', IEEE Transactions on Professional Communication, 43(3), p. 261.

Bowers, V. A. and Snyder, H. L. (1990) 'Concurrent versus retrospective verbal protocol for comparing window usability', in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 1270–1274. Available at: http://pro.sagepub.com/content/34/17/1270.short (Accessed: 31 August 2016).

Bowles, M. A. (2010) The think-aloud controversy in second language research. Routledge. Available at: https://books.google.co.uk/books?hl=en&lr=&id=u_KMAgAAQBAJ&oi=fnd&pg=PP1&dq=think+aloud+is+in+itself+sufficient+to+cause+reactivity+&ots=MfaHfVhzCC&sig=3WqfwjXqKv5OWUrdmyZd6gMwEek (Accessed: 6 July 2016).

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

Brooke, J. and others (1996) ‘SUS-A quick and dirty usability scale’, Usability evaluation in industry, 189(194), pp. 4–7.

Bruun, A. and Stage, J. (2015) ‘An empirical study of the effects of three think-aloud protocols on identification of usability problems’, in Human-Computer Interaction. Springer, pp. 159–176. Available at: http://link.springer.com/chapter/10.1007/978-3-319-22668-2_14 (Accessed: 9 February 2017).

Bruun, A. and Stage, J. (2015) ‘An empirical study of the effects of three think-aloud protocols on identification of usability problems’, in Human-Computer Interaction. Springer, pp. 159–176. Available at: http://link.springer.com/chapter/10.1007/978-3-319-22668-2_14 (Accessed: 9 February 2017).

Brannen, J., 2005. Mixing methods: The entry of qualitative and quantitative approaches into the research process. International journal of social research methodology, 8(3), pp.173-184.

BUTLER, K. A. (1996) interactions, 3(null), p. 59.

Chadha, R. et al. (2016) ‘Application of Data Mining Techniques on Heart Disease Prediction: A Survey’, in Emerging Research in Computing, Information, Communication and Applications. Springer, pp. 413–426. Available at: http://link.springer.com/chapter/10.1007/978-81-322-2553-9_38 (Accessed: 16 February 2017).

Chan, R. C. K., Hoosain, R. and Lee, T. M. C. (2002) ‘Talking While Performing a Task: A Better Attentional Performance inPatients With Closed Head Injury?’, Journal of Clinical & Experimental Neuropsychology, 24(5), p. 695.

Chatterjee, M., Grice, R. A. and Adali, S. (2003) ‘Applying usability engineering methodology to building a search agent for Web applications’, in Professional Communication Conference, 2003. IPCC 2003. Proceedings. IEEE International. IEEE, p. 8–pp. Available at: http://ieeexplore.ieee.org/abstract/document/1245481/ (Accessed: 13 February 2017).

Chattratichart, J. and Lindgaard, G. (2008) ‘A comparative evaluation of heuristic-based usability inspection methods’, in CHI’08 extended abstracts on Human factors in computing systems. ACM, pp. 2213–2220. Available at: http://dl.acm.org/citation.cfm?id=1358654 (Accessed: 10 February 2017).

Chen, W. et al. (2018) ‘Automated comprehensive evaluation approach for user interface satisfaction based on concurrent think-aloud method’, Universal Access in the Information Society, pp. 1–13. doi: 10.1007/s10209-018-0610-z.

Chin, J. M. and Schooler, J. W. (2008) ‘Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing’, European Journal of Cognitive Psychology, 20(3), pp. 396–413. doi: 10.1080/09541440701728623.

Chi, M. (1997). Quantifying qualitative analyses of verbal data: a practical guide. Journal of the Learning Sciences. 6 (3). pp. 271-315.

Clemmensen, T., (2012), “Usability problem identification in culturally diverse settings. *Information Systems Journal*, *22*(2), pp.151-175.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q. and Yammiyavar, P., (2009). Cultural cognition in usability evaluation. *Interacting with computers*, *21*(3), pp.212-220.

Cohen, A. D. (2000) ‘Exploring strategies in test taking: Fine-tuning verbal reports from respondents’, Learner-directed assessment in ESL, pp. 127–150.

Cooke, L. (2010) ‘Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach’, Professional Communication, IEEE Transactions on, 53(3), pp. 202–215.

Cotton, D. and Gresty, K. (2006) ‘Reflecting on the think-aloud method for evaluating e-learning’, British Journal of Educational Technology, 37(1), pp. 45–54.

Creswell, J.W., V.L. Plano Clark, M.L. Gutmann, and W.E. Hanson. 2003. Advanced mixed methods research designs. In Handbook of mixed methods in social and behavioral research, eds. A.Tashakkori and C. Teddlie, 209–240. Thousand Oaks, CA: Sage Publications.

Darin, T., Andrade, R. and Sánchez, J. (2018) ‘CLUE: A Usability Evaluation Checklist for Multimodal Video Game Field Studies with Children Who Are Blind’, in Proceedings of the 51st Hawaii International Conference on System Sciences.

Davies, E. L. et al. (2017) ‘Acceptability of targeting social embarrassment in a digital intervention to reduce student alcohol consumption: A qualitative think aloud study’, DIGITAL HEALTH, 3, p. 2055207617733405.

Denning, S. et al. (1990) ‘The value of thinking-aloud protocols in industry: A case study at Microsoft Corporation’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 1285–1289. Available at: http://pro.sagepub.com/content/34/17/1285.short (Accessed: 5 July 2016).

Denton, A. H., Moody, D. A. and Bennett, J. C. (2016) ‘Usability Testing as a Method to Refine a Health Sciences Library Website’, Medical Reference Services Quarterly, 35(1), pp. 1–15. doi: 10.1080/02763869.2016.1117280.

Dickson-Swift, V., James, E. L., & Liamputtong, P. (2008). Undertaking sensitive research in the health and social sciences: Managing boundaries, emotions and risks. Cambridge, NY: Cambridge University Press.

Dougherty, T., 2020. Informed consent, disclosure, and understanding. Philosophy & Public Affairs, 48(2), pp.119-150.

Du Plessis, J.-J. and Mwalemba, G. (2016) ‘Adoption of emerging technologies into ERP systems landscape: A South African study’, in Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), IEEE International Conference on. IEEE, pp. 395–399. Available at: http://ieeexplore.ieee.org/abstract/document/7737373/ (Accessed: 16 February 2017).

Dumas, J. S. and Fox, J. E. (2009) ‘Usability testing: Current practice and future directions’, Human-Computer Interaction: Development Process, 231. Available at: https://books.google.co.uk/books?hl=en&lr=&id=clMsHX-JfyMC&oi=fnd&pg=PA231&dq=Usability+Testing:+Current+Practice+and+Future+Directions.&ots=7rfR8vczGw&sig=xMrSASHLYJnO7gARVM75XD_29bA (Accessed: 6 March 2017).

Dumas, J. S. and Redish, J. (1999) A practical guide to usability testing. Intellect Books. Available at: https://books.google.co.uk/books?hl=en&lr=&id=4lge5k_F9EwC&oi=fnd&pg=PR5&dq=dumas+and+redish+1999&ots=vqiccDf8vD&sig=3x_52vAYdInOjKVLYcUWv3_xVk4 (Accessed: 23 June 2016).

Dumas, J. S., Molich, R. and Jeffries, R. (2004) ‘Describing Usability Problems: Are we sending the right message?’, interactions, 11(4), pp. 24–29.

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, NY, USA: ACM (DIS '00), pp. 289–296. doi: 10.1145/347642.347766.

Eccles, D. W. and Arsal, G. (2017) 'The think aloud method: what is it and how do I use it?', Qualitative Research in Sport, Exercise and Health, 9(4), pp. 514–531.

Edwards, H. et al. (2013) 'Reflecting on the Pret a Rapporter Framework via a Field Study of Adolescents' Perceptions of Technology and Exercise', Human Technology; Special issue on The End of Cognition, 9(2), pp. 131–156.

Eger, N. et al. (2007) 'Cueing retrospective verbal reports in usability testing through eye-movement replay', in Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1. British Computer Society, pp. 129–137. Available at: http://dl.acm.org/citation.cfm?id=1531312 (Accessed: 17 August 2016).

Elbabour, F., Alhadreti, O. and Mayhew, P. (2017) 'Eye Tracking in Retrospective Think-aloud Usability Testing: Is There Added Value?', J. Usability Studies, 12(3), pp. 95–110.

Elling, S., Lentz, L. and de Jong, M. (2011) 'Retrospective Think-aloud Method: Using Eye Movements As an Extra Cue for Participants' Verbalizations', in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI '11), pp. 1161–1170. doi: 10.1145/1978942.1979116.

Elling, S., Lentz, L. and de Jong, M. (2012) 'Combining Concurrent Think-Aloud Protocols and Eye-Tracking Observations: An Analysis of Verbalizations and Silences', IEEE Transactions on Professional Communication, 55(3), pp. 206–220. doi: 10.1109/TPC.2012.2206190.

Ellis, K., Quigley, M. and Power, M. (2008) 'Experiences in ethical usability testing with children', Journal of Information Technology Research (JITR), 1(3), pp. 1–13.

Endo, Y., MacKenzie, D. C. and Arkin, R. C. (2004) 'Usability evaluation of high-level user assistance for robot mission specification', IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(2), pp. 168–180.

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), pp. 981–987.

Ericsson, K. A. and Fox, M. C. (2011) 'Thinking aloud is not a form of introspection but a qualitatively different methodology: Reply to Schooler (2011)', Psychological Bulletin, 137(2), pp. 351–354. doi: 10.1037/a0022388.

Ericsson, K. A. and Simon, H. A. (1980) 'Verbal reports as data.', Psychological review, 87(3), p. 215.

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

Fan, M., Shi, S. and Truong, K.N., (2020) Practices and Challenges of Using Think-Aloud Protocols in Industry: An International Survey. *Journal of Usability Studies*, *15*(2).

Fernandez, A., Insfran, E. and Abrahão, S. (2011) 'Usability evaluation methods for the web: A systematic mapping study', Information and Software Technology, 53(8), pp. 789–817.

Flower, L. and Hayes, J. R. (1981) ‘A cognitive process theory of writing’, College composition and communication, 32(4), pp. 365–387.

Følstad, A. and Hornbæk, K. (2010) ‘Work-domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing’, Journal of Systems and Software. (Interplay between Usability Evaluation and Software Development), 83(11), pp. 2019–2030. doi: 10.1016/j.jss.2010.02.026.

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), p. 316.

Freeman, B. (2011) ‘Triggered Think-aloud Protocol: Using Eye Tracking to Improve Usability Test Moderation’, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI ’11), pp. 1171–1174. doi: 10.1145/1978942.1979117.

Frøkj\a er, E. and Hornb\a ek, K. (2005) ‘Cooperative usability testing: complementing usability tests with user-supported interpretation sessions’, in CHI’05 extended abstracts on Human factors in computing systems. ACM, pp. 1383–1386. Available at: http://dl.acm.org/citation.cfm?id=1056922 (Accessed: 10 February 2017).

Frøkj\a er, E., Hertzum, M. and Hornb\a ek, 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, NY, USA: ACM (CHI ’00), pp. 345–352. doi: 10.1145/332040.332455.

Furniss, D., Blandford, A. and Curzon, P. (2007) ‘Usability Evaluation Methods in Practice: Understanding the Context in Which They Are Embedded’, in Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore! New York, NY, USA: ACM (ECCE ’07), pp. 253–256. doi: 10.1145/1362550.1362602.

Gerjets, P., Kammerer, Y. and Werner, B. (2011) ‘Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data’, Learning and Instruction, 21(2), pp. 220–231.

Gill, A. M. and Nonnecke, B. (2012) ‘Think aloud: effects and validity’, in Proceedings of the 30th ACM international conference on Design of communication. ACM, pp. 31–36. Available at: http://dl.acm.org/citation.cfm?id=2379065 (Accessed: 13 February 2017).

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

GRAY, W. D. and SALZMAN, M. C. (1998a) Human - Computer Interaction, 13(null), p. 203.

Gray, W. D. and Salzman, M. C. (1998b) ‘Repairing Damaged Merchandise: A Rejoinder’, Human–Computer Interaction, 13(3), pp. 325–335. doi: 10.1207/s15327051hci1303_4.

GRAY, W. D. and SALZMAN, M. C. (1998b) Human - Computer Interaction, 13(null), p. 325.

Greenberg, S. and Buxton, B. (2008) ‘Usability Evaluation Considered Harmful (Some of the Time)’, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI ’08), pp. 111–120. doi: 10.1145/1357054.1357074.

Guan, Z. et al. (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. New York, NY, USA: ACM (CHI ’06), pp. 1253–1262. doi: 10.1145/1124772.1124961.

Gulliksen, J., Boivie, I. and Göransson, B. (2006) ‘Usability professionals—current practices and future development’, Interacting with computers, 18(4), pp. 568–600.

Guo, X. and Huang, L.-S. (2018) ‘Are L1 and L2 strategies transferable? An exploration of the L1 and L2 writing strategies of Chinese graduate students’, The Language Learning Journal, pp. 1–23.

Hackman, G. S. and Biers, D. W. (1992) ‘Team Usability Testing: Are two Heads Better than One?’, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 36(16), pp. 1205–1209. doi: 10.1177/154193129203601605.

Hasan, L., Morris, A. and Probets, S. (2012a) ‘A comparison of usability evaluation methods for evaluating e-commerce websites’, Behaviour & Information Technology, 31(7), pp. 707–737. doi: 10.1080/0144929X.2011.596996.

Hasan, L., Morris, A. and Probets, S. (2012b) ‘A comparison of usability evaluation methods for evaluating e-commerce websites’, Behaviour & Information Technology, 31(7), pp. 707–737. doi: 10.1080/0144929X.2011.596996.

Hassenzahl, M. (2000) ‘Prioritizing usability problems: Data-driven and judgement-driven severity estimates’, Behaviour & Information Technology, 19(1), pp. 29–42.

Held, J. E. and Biers, D. W. (1992a) ‘Software usability testing: Do evaluator intervention and task structure make any difference?’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications Sage CA: Los Angeles, CA, pp. 1215–1219. Available at: http://journals.sagepub.com/doi/abs/10.1177/154193129203601607 (Accessed: 10 February 2017).

Held, J. E. and Biers, D. W. (1992b) ‘Software Usability Testing: Do Evaluator Intervention and Task Structure Make any Difference?’, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 36(16), pp. 1215–1219. doi: 10.1177/154193129203601607.

Hertzum, M. (2016a) ‘A Usability Test is Not an Interview’, interactions, 23(2), pp. 82–84. doi: 10.1145/2875462.

Hertzum, M. (2016b) ‘Usability Testing: Too Early? Too Much Talking? Too Many Problems?’, Journal of Usability Studies, 11(3), pp. 83–88.

Hertzum, M. (2018) ‘Commentary: Usability—A Sensitizing Concept’, Human–Computer Interaction, 33(2), pp. 178–181.

Hertzum, M. and Clemmensen, T. (2012) ‘How do usability professionals construe usability?’, International Journal of Human-Computer Studies, 70(1), pp. 26–42. doi: 10.1016/j.ijhcs.2011.08.001.

Hertzum, M. and Holmegaard, K. D. (2013) ‘Thinking aloud in the presence of interruptions and time constraints’, International Journal of Human-Computer Interaction, 29(5), pp. 351–364.

Hertzum, M. and Holmegaard, K. D. (2015) ‘Thinking aloud influences perceived time’, Human Factors, 57(1), pp. 101–109.

Hertzum, M. and Jacobsen, N. E. (2001) ‘The evaluator effect: A chilling fact about usability evaluation methods’, International Journal of Human-Computer Interaction, 13(4), pp. 421–443.

Hertzum, M. et al. (2011) ‘Personal Usability Constructs: How People Construe Usability Across Nationalities and Stakeholder Groups’, International Journal of Human–Computer Interaction, 27(8), pp. 729–761. doi: 10.1080/10447318.2011.555306.

Hertzum, M., Borlund, P. and Kristoffersen, K. B. (2015) ‘What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions’, International Journal of Human-Computer Interaction, 31(9), pp. 557–570.

Hertzum, M., Hansen, K. D. and Andersen, H. H. (2009) ‘Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload?’, Behaviour & Information Technology, 28(2), pp. 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 & Information Technology, 33(2), pp. 144–162. doi: 10.1080/0144929X.2013.783114.

Hoegh, R. T. et al. (2006) ‘The Impact of Usability Reports and User Test Observations on Developers’ Understanding of Usability Data: An Exploratory Study’, International Journal of Human-Computer Interaction, 21(2), pp. 173–196. doi: 10.1207/s15327590ijhc2102_4.

Hofer, B. K. (2004) ‘Epistemological understanding as a metacognitive process: Thinking aloud during online searching’, Educational Psychologist, 39(1), pp. 43–55.

Hogan, K. (1999) ‘Thinking Aloud Together: A Test of an Intervention To Foster Students’ Collaborative Scientific Reasoning.’, Journal of Research in Science Teaching, 36(10), pp. 1085–1109.

HOLLERAN, P. A. (1991) ‘A methodological note on pitfalls in usability testing’, Behaviour & Information Technology, 10(5), pp. 345–357. doi: 10.1080/01449299108924295.

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

Holzinger, A., Brugger, M. and Slany, W. (2011) ‘Applying aspect oriented programming in usability engineering processes: On the example of tracking usage information for remote usability testing’, in e-Business (ICE-B), 2011 Proceedings of the International Conference on. IEEE, pp. 1–4. Available at: http://ieeexplore.ieee.org/abstract/document/6731091/ (Accessed: 13 February 2017).

Hoppmann, T. K. (2009) ‘Examining the “point of frustration”. The think-aloud method applied to online search tasks’, Quality & Quantity, 43(2), pp. 211–224.

Hornb\a ek, K. and Frøkj\a er, E. (2008) ‘A study of the evaluator effect in usability testing’, Human–Computer Interaction, 23(3), pp. 251–277.

Hornbæk, K. (2010) ‘Dogmas in the assessment of usability evaluation methods’, Behaviour & Information Technology, 29(1), pp. 97–111. doi: 10.1080/01449290801939400.

Howarth, J., Andre, T. S. and Hartson, R. (2007) ‘A Structured Process for Transforming Usability Data into Usability Information’, J. Usability Studies, 3(1), pp. 7–23.

Huh, J. et al. (2007) ‘Beyond Usability: Taking Social, Situational, Cultural, and Other Contextual Factors into Account’, in CHI ’07 Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI EA ’07), pp. 2113–2116. doi: 10.1145/1240866.1240961.

Hundt, A. S., Adams, J. A. and Carayon, P. (2016) ‘A Collaborative Usability Evaluation (CUE) Model for Health IT Design and Implementation’, International Journal of Human–Computer Interaction, 0(0), pp. 1–11. doi: 10.1080/10447318.2016.1263430.

ISO 9241-11, S. (1998) ‘9241-11. 1998’, Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)–Part II Guidance on Usability.

ISO 9241-210, I. (2010) ‘9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems’, International Standardization Organization (ISO). Switzerland.

Israel, M. (2015). Research ethics and integrity for social scientists. London, England: Sage.

Jacob, E. et al. (2012) ‘Usability testing of a Smartphone for accessing a web-based e-diary for self-monitoring of pain and symptoms in sickle cell disease’, Journal of pediatric hematology/oncology, 34(5), p. 326.

Jacobsen, N. E., Hertzum, M. and John, B. E. (1998) ‘The evaluator effect in usability studies: Problem detection and severity judgments’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 1336–1340. Available at: http://pro.sagepub.com/content/42/19/1336.short (Accessed: 10 February 2017).

Jensen, J. J. (2007) ‘Evaluating in a Healthcare Setting: A Comparison Between Concurrent and Retrospective Verbalisation’, in Proceedings of the 12th International Conference on Human-computer Interaction: Interaction Design and Usability. Berlin, Heidelberg: Springer-Verlag (HCI’07), pp. 508–516. Available at: http://dl.acm.org/citation.cfm?id=1772490.1772548 (Accessed: 12 May 2016).

Jewell, C. and Salvetti, F. (2012) ‘Towards a Combined Method of Web Usability Testing: An Assessment of the Complementary Advantages of Lab Testing, Pre-session Assignments, and Online Usability Services’, in CHI ’12 Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI EA ’12), pp. 1865–1870. doi: 10.1145/2212776.2223720.

Jibb, L. et al. (2012) ‘Pain Squad: usability testing of a multidimensional electronic pain diary for adolescents with cancer’, The Journal of Pain, 13(4), p. S23.

JØRGENSEN, A. H. (1990) ‘Thinking-aloud in user interface design: a method promoting cognitive ergonomics’, Ergonomics, 33(4), pp. 501–507. doi: 10.1080/00140139008927157.

Karani, A., Thanki, H. and Achuthan, S., 2021. Impact of university website usability on satisfaction: a structural equation modelling approach. Management and Labour Studies, 46(2), pp.119-138.

Karsh, B. T. (2004) ‘Beyond usability: designing effective technology implementation systems to promote patient safety’, Quality and Safety in Health Care, 13(5), pp. 388–394.

Kawalek, J., Stark, A. and Riebeck, M. (2008) ‘A new approach to analyze human-mobile computer interaction’, Journal of usability studies, 3(2), pp. 90–98.

Kennedy, S. (1989) ‘Using Video in the BNR Usability Lab’, SIGCHI Bull., 21(2), pp. 92–95. doi: 10.1145/70609.70624.

Kortum, P. and Peres, S. C. (2014) ‘The Relationship Between System Effectiveness and Subjective Usability Scores Using the System Usability Scale’, International Journal of Human–Computer Interaction, 30(7), pp. 575–584. doi: 10.1080/10447318.2014.904177.

Koscielny, J. et al. (2005) ‘Use of the platelet reactivity index by Grotemeyer, platelet function analyzer, and retention test Homburg to monitor therapy with antiplatelet drugs’, in Seminars in thrombosis and hemostasis. Copyright\copyright 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA., pp. 464–469. Available at: https://www.thieme-connect.com/products/ejournals/html/10.1055/s-2005-916682 (Accessed: 13 February 2017).

Krahmer (2004) ‘Thinking About Thinking Aloud: A Comparison of TwoVerbal Protocols for Usability Testing’, IEEE Transactions on Professional Communication, 47(2), pp. 105–117. doi: 10.1109/TPC.2004.828205.

Krahmer, E. and Ummelen, N. (2004) ‘Thinking about thinking aloud: A comparison of two verbal protocols for usability testing’, IEEE Transactions on Professional Communication, 47(2), pp. 105–117.

Kumar, V. (2017) ‘The think aloud method: Some concerns addressed’, Journal of Modern Languages, 15(1), pp. 13–25.

Kurosu, M. et al. (2004) ‘Trends in Usability Research and Activities in Japan’, International Journal of Human-Computer Interaction, 17(1), pp. 103–124. doi: 10.1207/s15327590ijhc1701_8.

Lalloo, C. et al. (2013) ‘Adapting the Iconic Pain Assessment Tool Version 2 (IPAT2) for adults and adolescents with arthritis pain through usability testing and refinement of pain quality icons’, The Clinical journal of pain, 29(3), pp. 253–264.

Lavery, D., Cockton, G. and Atkinson, M. P. (1997) ‘Comparison of evaluation methods using structured usability problem reports’, Behaviour & Information Technology, 16(4–5), pp. 246–266. doi: 10.1080/014492997119824.

Law, E. L.-C. and Hvannberg, E. T. (2008) ‘Consolidating Usability Problems with Novice Evaluators’, in Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges. New York, NY, USA: ACM (NordiCHI ’08), pp. 495–498. doi: 10.1145/1463160.1463228.

Lee, Y. K. et al. (2018) ‘Usability and utility evaluation of the web-based “Should I Start Insulin?” patient decision aid for patients with type 2 diabetes among older people’, Informatics for Health and Social Care, 43(1), pp. 73–83.

Lentz, L. and Elling, S. (2014) ‘User Page Reviews in Usability Testing’, Evaluating Websites and Web Services: Interdisciplinary Perspectives on User Satisfaction: Interdisciplinary Perspectives on User Satisfaction, p. 95.

Lewis, J. R. (2014) ‘Usability: Lessons Learned … and Yet to Be Learned’, International Journal of Human-Computer Interaction, 30(9), pp. 663–684. doi: 10.1080/10447318.2014.930311.

Liapis, A., Katsanos, C. and Xenos, M. (2018) 'Don't Leave Me Alone: Retrospective Think Aloud supported by Real-time Monitoring of Participant's Physiology', arXiv preprint arXiv:1802.04090.

Lin, H. X., Choong, Y.-Y. and Salvendy, G. (1997) 'A proposed index of usability: A method for comparing the relative usability of different software systems', Behaviour & Information Technology, 16(4–5), pp. 267–277. doi: 10.1080/014492997119833.

Lindgaard, G. and Chattratichart, J. (2007) 'Usability testing: what have we overlooked?', in Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp. 1415–1424. Available at: http://dl.acm.org/citation.cfm?id=1240839 (Accessed: 10 February 2017).

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 & Management, 47(3), pp. 336–348. doi: 10.1016/j.ipm.2010.08.001.

Marti, P. and Rizzo, A. (2003) 'Levels of design: from usability to experience', in HCI International 2003, 10th International Conference on Human-Computer Interaction. Available at: https://www.researchgate.net/profile/Patrizia_Marti/publication/228584405_Levels_of_design_from_usability_to_experience/links/0912f51320c8b3759e000000.pdf (Accessed: 13 February 2017).

McDonald, S. and Petrie, H. (2013) 'The Effect of Global Instructions on Think-aloud Testing', in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI '13), pp. 2941–2944. doi: 10.1145/2470654.2481407.

McDonald, S., Edwards, H. M. and Zhao, T. (2012) 'Exploring think-alouds in usability testing: An international survey', Professional Communication, IEEE Transactions on, 55(1), pp. 2–19.

McDonald, S., McGarry, K. and Willis, L. M. (2013) 'Thinking-aloud about web navigation the relationship between think-aloud instructions, task difficulty and performance', in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 2037–2041. Available at: http://pro.sagepub.com/content/57/1/2037.short (Accessed: 5 May 2016).

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. ACM, pp. 437–440. Available at: http://dl.acm.org/citation.cfm?id=1182531 (Accessed: 13 February 2017).

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), pp. 647–660.

McDonald, S., Zhao, T. and Edwards, H. M. (2015) 'Look Who's Talking: Evaluating the Utility of Interventions During an Interactive Think-Aloud', Interacting with Computers, p. iwv014.

Meissner, C. A. and Brigham, J. C. (2001) 'A meta-analysis of the verbal overshadowing effect in face identification', Applied Cognitive Psychology, 15(6), pp. 603–616.

Mentis, H. and Gay, G. (2003) 'User recalled occurrences of usability errors: Implications on the user experience', in CHI'03 extended abstracts on Human factors in computing systems.

ACM, pp. 736–737. Available at: http://dl.acm.org/citation.cfm?id=765959 (Accessed: 13 February 2017).

Molich, R. et al. (2004) ‘Comparative usability evaluation’, Behaviour & Information Technology, 23(1), pp. 65–74.

Molich, R., Jeffries, R. and Dumas, J. S. (2007) ‘Making usability recommendations useful and usable’, Journal of Usability Studies, 2(4), pp. 162–179.

Morais, F. de, Schaab, B. and Jaques, P. (2017) ‘The think aloud method for qualitative evaluation of an intelligent tutoring system interface’, in 2017 Twelfth Latin American Conference on Learning Technologies (LACLO). 2017 Twelfth Latin American Conference on Learning Technologies (LACLO), pp. 1–8. doi: 10.1109/LACLO.2017.8120904.

Myers, B. (1994) ‘Challenges of HCI design and implementation’, interactions, 1(1), pp. 73–83.

Myers, B. A. (1993) Why are human-computer interfaces difficult to design and implement. DTIC Document. Available at: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA268843 (Accessed: 13 February 2017).

N⊘rgaard, M. and Hornbæk, K. (2009) ‘Exploring the Value of Usability Feedback Formats’, International Journal of Human–Computer Interaction, 25(1), pp. 49–74. doi: 10.1080/10447310802546708.

Natesan, D., Walker, M. and Clark, S. (2016) ‘Cognitive Bias in Usability Testing’, in Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care. SAGE Publications, pp. 86–88. Available at: http://hcs.sagepub.com/content/5/1/86.short (Accessed: 13 February 2017).

Nawaz, A. and Clemmensen, T. (2013) ‘Website Usability in Asia “From Within”: An Overview of a Decade of Literature’, International Journal of Human–Computer Interaction, 29(4), pp. 256–273. doi: 10.1080/10447318.2013.765764.

Nielsen, J. (1994) Usability engineering. Elsevier. Available at: https://books.google.co.uk/books?hl=en&lr=&id=DBOowF7LqIQC&oi=fnd&pg=PP1&dq=nielsen+1993+usability+engineering&ots=Bk68WQIVyP&sig=hP1fhnokBfdvYdlB6P6spUs1Wf8 (Accessed: 23 June 2016).

Nisbett, R. E. and Wilson, T. D. (1977) ‘Telling more than we can know: Verbal reports on mental processes.’, Psychological review, 84(3), p. 231.

Nørgaard, M. and Hornb\a ek, K. (2008) ‘Working together to improve usability: challenges and best practices’, in. Copenhagen University Technical Report. Available at: http://www.diku.dk/forskning/Publikationer/tekniske_rapporter/2008/08-03.pdf (Accessed: 10 February 2017).

Nørgaard, M. and Hornbaek, K. (2006) ‘What Do Usability Evaluators Do in Practice?: An Explorative Study of Think-aloud Testing’, in Proceedings of the 6th Conference on Designing Interactive Systems. New York, NY, USA: ACM (DIS ’06), pp. 209–218. doi: 10.1145/1142405.1142439.

Ohnemus, K. R. and Biers, D. W. (1993) ‘Retrospective versus Concurrent Thinking-Out-Loud in Usability Testing’, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 37(17), pp. 1127–1131. doi: 10.1177/154193129303701701.

Oliveira, A. et al. (2013) ‘Usability testing of a respiratory interface using computer screen and facial expressions videos’, Computers in biology and medicine, 43(12), pp. 2205–2213.

Olmsted-Hawala, E. L. et al. (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, NY, USA: ACM (CHI ’10), pp. 2381–2390. doi: 10.1145/1753326.1753685.

Olsen, A., Smolentzov, L. and Strandvall, T. (2010) ‘Comparing Different Eye Tracking Cues when Using the Retrospective Think Aloud Method in Usability Testing’, in Proceedings of the 24th BCS Interaction Specialist Group Conference. Swinton, UK, UK: British Computer Society (BCS ’10), pp. 45–53. Available at: http://dl.acm.org/citation.cfm?id=2146303.2146310 (Accessed: 12 May 2016).

Page, C. and Rahimi, M. (1995) ‘Concurrent and retrospective verbal protocols in usability testing: Is there value added in collecting both?’, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 223–227. Available at: http://pro.sagepub.com/content/39/4/223.short (Accessed: 10 February 2017).

Pätsch, G., Mandl, T. and Womser-Hacker, C. (2014) ‘Using Sensor Graphs to Stimulate Recall in Retrospective Think-aloud Protocols’, in Proceedings of the 5th Information Interaction in Context Symposium. New York, NY, USA: ACM (IIiX ’14), pp. 303–307. doi: 10.1145/2637002.2637048.

Paz, F. and Pow-Sang, J. A. (2016) ‘A systematic mapping review of usability evaluation methods for software development process’, International Journal of Software Engineering and Its Applications, 10(1), pp. 165–178.

Pepper, D. et al. (2018) ‘Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics’, International Journal of Research & Method in Education, 41(1), pp. 3–16.

Peute, L. W. P., de Keizer, N. F. and Jaspers, M. W. M. (2015) ‘The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool’, Journal of Biomedical Informatics, 55, pp. 1–10. doi: 10.1016/j.jbi.2015.02.006.

Pitkänen, J., Pitkäranta, M. and Nieminen, M. (2012) ‘Usability Testing in Real Context of Use: The User-triggered Usability Testing’, in Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design. New York, NY, USA: ACM (NordiCHI ’12), pp. 797–798. doi: 10.1145/2399016.2399153.

Poole, A. and Ball, L. J. (no date) ‘Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects’. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.5691&rep=rep1&type=pdf (Accessed: 3 May 2017).

Potosnak, K., 1988. Human factors-recipe for a usability test. IEEE Software, 5(6), pp.83-84.

Pressley, M. and Afflerbach, P. (1995) Verbal protocols of reading: The nature of constructively responsive reading. Routledge.

Ramey, J. et al. (2006) ‘Does Think Aloud Work?: How Do We Know?’, in CHI ’06 Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: ACM (CHI EA ’06), pp. 45–48. doi: 10.1145/1125451.1125464.

Rashid, S. et al. (2013) ‘Preliminary Usability Testing with Eye Tracking and FCAT Analysis on Occupational Safety and Health Websites’, Procedia - Social and Behavioral Sciences. (The 9th International Conference on Cognitive Science), 97, pp. 737–744. doi: 10.1016/j.sbspro.2013.10.295.

Redish, J. and Dumas, J. (1999) ‘A practical guide to usability testing’, Intellect books.

Resnik, D.B., 2018. The ethics of research with human subjects: Protecting people, advancing science, promoting trust (Vol. 74). Springer.

Resnik, D.B., 2021. Informed consent, understanding, and trust. The American Journal of Bioethics, 21(5), pp.61-63.

Research ethics – The University of Sunderland. Available from: https://www.sunderland.ac.uk/more/research/research-governance-integrity/ethics/ [Accessed: 27th April 2023]

Ritthiron, S. and Jiamsanguanwong, A. (2017) ‘Usability Evaluation of the University Library Network’s Website Using an Eye-Tracking Device’, in Proceedings of the International Conference on Advances in Image Processing. New York, NY, USA: ACM (ICAIP 2017), pp. 184–188. doi: 10.1145/3133264.3133294.

Rodríguez-Pose, A. and Hardy, D. (2015) ‘Addressing poverty and inequality in the rural economy from a global perspective’, Applied Geography, 61, pp. 11–23. doi: 10.1016/j.apgeog.2015.02.005.

Røsand, T. (2012) Think Aloud Methods with Eye Tracking in Usability Testing: A comparison study with different task types. Institutt for datateknikk og informasjonsvitenskap. Available at: https://brage.bibsys.no/xmlui/handle/11250/253128 (Accessed: 13 February 2017).

Røsand, T. (no date) ‘Think Aloud Methods with Eye Tracking in Usability Testing’. Available at: http://daim.idi.ntnu.no/masteroppgaver/006/6309/masteroppgave.pdf (Accessed: 13 February 2017).

Russ, A. L. and Saleem, J. J. (2018) ‘Ten Factors to Consider when Developing Usability Scenarios and Tasks for Health Information Technology’, Journal of biomedical informatics.

Russo, J. E., Johnson, E. J. and Stephens, D. L. (1989) ‘The validity of verbal protocols’, Memory & cognition, 17(6), pp. 759–769.

Schellings, G. L. et al. (2013) ‘Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols’, European journal of psychology of education, 28(3), pp. 963–990.

Schneider, J. F. and Reichl, C. (2006) ‘Exploring ease in thinking aloud’, Psychological reports, 98(1), pp. 85–90.

Seffah, A. and Habieb-Mammar, H. (2009) ‘Usability engineering laboratories: limitations and challenges toward a unifying tools/practices environment’, Behaviour & Information Technology, 28(3), pp. 281–291. doi: 10.1080/01449290701803482.

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. New York, NY, USA: ACM (NordiCHI ’08), pp. 344–352. doi: 10.1145/1463160.1463198.

Silva, J. L., Campos, J. C. and Paiva, A. C. (2008) 'Model-based user interface testing with Spec Explorer and ConcurTaskTrees', Electronic Notes in Theoretical Computer Science, 208, pp. 77–93.

Sivaji, A., Nielsen, S. F. and Clemmensen, T. (2016) 'A Textual Feedback Tool for Empowering Participants in Usability and UX Evaluations', International Journal of Human–Computer Interaction, 0(0), pp. 1–14. doi: 10.1080/10447318.2016.1243928.

Smith, A., 2011. Issues in adapting usability testing for global usability. In *Global usability* (pp. 23-38). Springer, London.

Smith, H., Fitzpatrick, G. and Rogers, Y. (2004) 'Eliciting reactive and reflective feedback for a social communication tool: a multi-session approach', in Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques. ACM, pp. 39–48. Available at: http://dl.acm.org/citation.cfm?id=1013123 (Accessed: 13 February 2017).

Spool, J. M. (1999) Web site usability: a designer's guide. Morgan Kaufmann. Available at: https://books.google.co.uk/books?hl=en&lr=&id=zPl8e4W4dvMC&oi=fnd&pg=PR13&dq=J.+Spool+et+al+Web+Site+Usability+a+designers+guide&ots=3gl4qVJAuW&sig=sjWSUjLWNKquK3_C2HJ3Qj3N254 (Accessed: 2 March 2017).

Srinivas, P., Cornet, V. and Holden, R. (2016) 'Human Factors Analysis, Design, and Evaluation of Engage, a Consumer Health IT Application for Geriatric Heart Failure Self-Care', International Journal of Human–Computer Interaction, 0(0), pp. 1–15. doi: 10.1080/10447318.2016.1265784.

Stefano, F., Borsci, S. and Stamerra, G. (2010) 'Web usability evaluation with screen reader users: implementation of the partial concurrent thinking aloud technique', Cognitive processing, 11(3), pp. 263–272.

Sutcliffe, A. and Hart, J. (2017) 'Analyzing the Role of Interactivity in User Experience', International Journal of Human–Computer Interaction, 33(3), pp. 229–240. doi: 10.1080/10447318.2016.1239797.

SZCZUR, M. (1994) 'Usability testing — on a budget: a NASA usability test case study', Behaviour & Information Technology, 13(1–2), pp. 106–118. doi: 10.1080/01449299408914589.

Taylor, K. L. and Dionne, J.-P. (2000) 'Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing.', Journal of Educational Psychology, 92(3), pp. 413–425.

Taylor, S. et al. (2017) 'Usability testing of an electronic pain monitoring system for palliative cancer patients: A think-aloud study', Health informatics journal, p. 1460458217741754.

Todhunter, F. (2015) 'Using concurrent think-aloud and protocol analysis to explore student nurses' social learning information communication technology knowledge and skill development', Nurse Education Today, 35(6), pp. 815–822. doi: 10.1016/j.nedt.2015.01.010.

Tohidi, M. et al. (2006) 'User sketches: a quick, inexpensive, and effective way to elicit more reflective user feedback', in Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles. ACM, pp. 105–114. Available at: http://dl.acm.org/citation.cfm?id=1182487 (Accessed: 13 February 2017).

van den Haak, M. and van Hooijdonk, C. (2010) 'Evaluating consumer health information websites: The importance of collecting observational, user-driven data', in Professional

Communication Conference (IPCC), 2010 IEEE International. IEEE, pp. 333–338. Available at: http://ieeexplore.ieee.org/abstract/document/5530031/ (Accessed: 13 February 2017).

Van den Haak, M. J. and de Jong, M. D. (2003) 'Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols', in Professional Communication Conference, 2003. IPCC 2003. Proceedings. IEEE International. IEEE, p. 3–pp. Available at: http://ieeexplore.ieee.org/abstract/document/1245501/ (Accessed: 13 February 2017).

Van den Haak, M. J., de Jong, M. D. and Schellens, P. J. (2004a) 'Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison', Interacting with computers, 16(6), pp. 1153–1170.

Van den Haak, M. J., de Jong, M. D. and Schellens, P. J. (2004b) 'Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison', Interacting with computers, 16(6), pp. 1153–1170.

Van den Haak, M. J., De Jong, M. D. and Schellens, P. J. (2007) 'Evaluation of an informational web site: three variants of the think-aloud method compared', Technical Communication, 54(1), pp. 58–71.

Van den Haak, M. J., de Jong, M. D. and Schellens, P. J. (2009) 'Evaluating municipal websites: A methodological comparison of three think-aloud variants', Government Information Quarterly, 26(1), pp. 193–202.

Van Den Haak, M., De Jong, M. and Jan Schellens, P. (2003) 'Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue', Behaviour & information technology, 22(5), pp. 339–351.

Van Waes, L. (2000) 'Thinking aloud as a method for testing the usability of websites: the influence of task variation on the evaluation of hypertext', IEEE transactions on professional communication, 43(3), pp. 279–291.

Vatrapu, R. and Pérez-Quiñones, M. A. (2006) 'Culture and usability evaluation: The effects of culture in structured interviews', Journal of usability studies, 1(4), pp. 156–170.

Velsen, L. van, Geest, T. van der and Klaassen, R. (2011) 'Identifying Usability Issues for Personalization During Formative Evaluations: A Comparison of Three Methods', International Journal of Human-Computer Interaction, 27(7), pp. 670–698. doi: 10.1080/10447318.2011.555304.

Waes, L. van (2000) 'Thinking aloud as a method for testing the usability of Websites: the influence of task variation on the evaluation of hypertext', IEEE Transactions on Professional Communication, 43(3), pp. 279–291. doi: 10.1109/47.867944.

Ward, J. L. and Hiller, S. (2005) 'Usability Testing, Interface Design, and Portals', Journal of Library Administration, 43(1–2), pp. 155–171. doi: 10.1300/J111v43n01_10.

Ward, R. D. and Marsden, P. H. (2004) 'Affective computing: problems, reactions and intentions', Interacting with Computers, 16(4), pp. 707–713.

Wells, A. T. (2006) 'Usability: reconciling theory and practice', in Proceedings of the 24th annual ACM international conference on Design of communication. ACM, pp. 99–104. Available at: http://dl.acm.org/citation.cfm?id=1166348 (Accessed: 13 February 2017).

White, P. (1980) ‘Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem.’ Available at: http://psycnet.apa.org/journals/rev/87/1/105/ (Accessed: 10 February 2017).

Whitehead, A. E. et al. (2018) ‘Investigating the relationship between cognitions, pacing strategies and performance in 16.1 km cycling time trials using a think aloud protocol’, Psychology of Sport and Exercise, 34, pp. 95–109.

Wichansky, A. M. (2000) ‘Usability testing in 2000 and beyond’, Ergonomics, 43(7), pp. 998–1006. doi: 10.1080/001401300409170.

Willis, L. M. and McDonald, S. (2016) ‘Retrospective protocols in usability testing: a comparison of Post-session RTA versus Post-task RTA reports’, Behaviour & Information Technology, pp. 1–16.

Wixon, D. (2003) ‘Evaluating Usability Methods: Why the Current Literature Fails the Practitioner’, interactions, 10(4), pp. 28–34. doi: 10.1145/838830.838870.

Wright, R. B. and Converse, S. A. (1992) ‘Method Bias and Concurrent Verbal Protocol in Software Usability Testing’, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 36(16), pp. 1220–1224. doi: 10.1177/154193129203601608.

Yanco, H. A., Drury, J. L. and Scholtz, J. (2004) ‘Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition’, Human–Computer Interaction, 19(1–2), pp. 117–149. doi: 10.1080/07370024.2004.9667342.

Yang, C., Hu, G. and Zhang, L. J. (2014) ‘Reactivity of concurrent verbal reporting in second language writing’, Journal of Second Language Writing, 24, pp. 51–70. doi: 10.1016/j.jslw.2014.03.002.

Yang, S. C. (2003) ‘Reconceptualizing think-aloud methodology: refining the encoding and categorizing techniques via contextualized perspectives’, Computers in Human Behavior, 19(1), pp. 95–115.

Yusop, N. S. M., Grundy, J. and Vasa, R. (2017) ‘Reporting Usability Defects: A Systematic Literature Review’, IEEE Transactions on Software Engineering, 43(9), pp. 848–867.

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, NY, USA: ACM (NordiCHI ’10), pp. 581–590. doi: 10.1145/1868914.1868979.

Zhao, T., McDonald, S. and Edwards, H. M. (2014) ‘The impact of two different think-aloud instructions in a usability test: a case of just following orders?’, Behaviour & Information Technology, 33(2), pp. 163–183.

ZIRKLER, D. and BALLMAN, D. R. (1994) ‘Usability testing in a competive market: lessons learned’, Behaviour & Information Technology, 13(1–2), pp. 191–197. doi: 10.1080/01449299408914598.

**APPENDICES**

## Appendix A: Materials from The Impact of Task-types on the Reactivity of the Classic Think-Aloud Study

### A1: Participant information sheet

# Participant Information Sheet

**Title:** A usability evaluation of Thomas Cook Holiday website (www.thomascook.com)

*Before you decide to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully.*

**Purpose of the study**

To evaluate Thomas Cook holiday websites in terms of its user-friendliness.

**Background and aim of the study**

The study aims to conduct a usability evaluation of the Thomas Cook website. The study is concerned with testing the user friendliness of a website; it is not a test of your ability to use the site. During the test you will complete a number of every day tasks with the product. These might involve searching for information or going through the process of choosing a holiday.

**Why have you been asked to take part?**

You have been asked because you are a representative user for the test product (Thomas cook holiday travel website). This means that you may use websites to book hotels or travel arrangements.

**What will happen to me if I take part?**

The study will take place at the usability Laboratory of the University of Sunderland, and will involve you and the author. As you interact with the website the research will record a video of your computer screen and audio data. He may also record your eye movements as you look at the product. The eye tracking equipment looks just like a standard PC monitor and uses infra-red sensors to track where your eyes are moving. It is just like using a normal monitor. During the test you may be asked think-aloud and complete some questions about the tasks you have completed e.g. whether you found them difficult. The session will last about 1 hour.

**Are there possible disadvantages and/or risks in taking part?**

There are no reasonably foreseeable discomforts, disadvantages, and risks associated with participating in this study.

**What are the possible benefits of taking part?**

Your participation will contribute to knowledge by providing usability practitioners with a deeper insight into the validity and reliability of their test data during usability testing.

**Will my taking part in this project be kept confidential?**

All data will be identified only by a participant code we will not maintain a record of your name. Data files will be kept for a period of two years and will be stored in a secure computer that is password protected.

**What will happen to the results of the research project?**

Results will be presented at conferences and written up in journals. Results are normally presented in terms of groups of individuals. The data collected during the course of this study might be used for additional or subsequent research if any individual data are presented; the data will be totally anonymous, without any means of identifying the individuals involved.

**Who is organising and funding the research?**

This research is self-sponsored to fulfill the requirement of a PhD

**Ethical review of the study**

The project has received ethical approval from the Research Ethics Committee (REC) of the University of Sunderland and it is being conducted in accordance with the University's Research Ethics Principles, Professional Codes of Practice and the law.

**Withdrawal of Participation**

Participating in this study is entirely voluntary and that refusal or withdrawal will involve no penalty or loss, now or in the future. If you decide at any time during the experiment that you no longer wish to participate, you may withdraw your consent without prejudice and the research will delete any data files in your presence.

**Contact for further information**

You may ask more questions about the study at any time. Please contact Obruche E. Orugbo via email: bg35nr@research.sunderland.ac.uk or on mobile: 07438341102

**A2: Participant Informed Consent Form**

# Research Informed Consent Form

Title of Project: An Investigation of the impact of Task Type on the Utility of Different Think-Aloud Approaches within Usability Testing

Investigator: Obruche Orugbo Emueakporavwa

Author email: bg35nr@research.sunderland.ac.uk

*Please read the following statements and, if you agree, tick the corresponding box to confirm agreement:*

| | |
|---|---|
| I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. | ☐ |
| I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. | ☐ |
| I understand that my data will be treated confidentially and any publication resulting from this work will report only data that does **not** identify me. | ☐ |
| I freely agree to participate in this study. | ☐ |

Signatures:

| | | |
|---|---|---|
| _____ | _____ | _____ |
| Name of participant (block capitals) | Date | Signature |
| _____ | _____ | _____ |
| Author (block capitals) | Date | Signature |

If you would like a copy of this consent form to keep, please ask the author. If you have any complaints or concerns about this research, you can direct these, in writing, to the Research supervisor Dr. Sharon McDonald by email at: sharon.mcdonald@sunderland.ac.uk

**A3: User Profile Questionnaire**

# User Profile Questionnaire

*Please fill, tick and circle where appropriate*

**Demography**

i. Participant's Number:11

ii. Gender: Male ☐ Female ☐

iii. Age:……...

iv. Nationality:.....

v. Highest educational qualification:

| | |
|---|---|
| PhD | ☐ |
| MSc | ☐ |
| BSc | ☐ |
| Senior Secondary | ☐ |

vi. Occupation:.....

**Profile Questions**

1. How long have you been using the internet?.....

2. How often do you use the internet?

| | |
|---|---|
| Daily | ☐ |
| Several times a week | ☐ |
| Several times a month | ☐ |
| Once a month | ☐ |

5. Have you used a travel website to look up different holiday destination in the past? Yes/No

If yes, when was the last time?..... and what is the name of the holiday site or website address:.....

6. Have you used Thomas Cook Travel website before? Yes/No

7. Have you ever participated in usability testing? Yes/No

**A4: Tasks Sets for: Study 1, An Investigation for the Impact of Task-types on the Reactivity of The Concurrent Think-aloud in Usability Testing**

# ACTIVITY 1

**Home > Holiday > Weddings and Honeymoons**
You are planning your forthcoming wedding abroad and have decided to use Thomas Cook. Please find out the phone number of their weddings team.

**Please write down the phone number you have found below**

**…….**

# ACTIVITY 2

**Home > Store locator > Store locator page**
You need to ring a local Thomas Cook travel store to enquire for certain information and visit the store for further enquiries. Please find the postcode of Thomas Cook store in address that is in North Shields

**Please write down the postcode below**

**…**

# ACTIVITY 3

**Home > Luxury > Sri-Lanka**
**Home > Luxury >** Cambodia
**Home > Luxury >** Thailand

You have decided to go on a holiday to visit either Cambodia, Sri-Lanka or Thailand.

Your decision to choose where to go depends on visa-free entry for British passport holder. Based on the information on Thomas Cook website which country would you visit.

**Please write down the country you would visit that has visa-free entry for British passport holder**

**……...**

# ACTIVITY 4

**Home > Footer > Download brochure > Brochure Store**

Your best friend has decided to visit Cyprus, she asked you to download an electronic brochure to help her find out about some of the resorts.

Please find a downloadable link to the electronic brochure that will assist your friend to find out more about Cyprus. Do not use search menu

**Please write the name of the brochure below**

**……**

# ACTIVITY 5

**Home > Credit card fee > Pricing terms and conditions**
You want to book a holiday and pay for it with your Barclaycard. Find out if Thomas Cook make a charge for this

**Credit card charge……**

# ACTIVITY 6

**Home > Flights > Available dates page > Basket > Passengers Information**
Please **DO NOT** use the search on the home page, use the navigation menu.

You wish to book a flight from London, Heathrow to Brussels, Belgium on 3rd October 2017 to return 17th October 2017.
You wish to use British Airways

**Please first complete the above task then include the following extras on your booking**

**Extras**

i. You would like to add a hotel accommodation offered by Thomas Cook that is close to Brussels City Centre
ii. You would like to add airport transfers for £200 or less
iii. Your total basket price should not be more than £2000

**Stop where you are required to enter passenger details**

**Please tick appropriately where you met the criteria**

You use British Airways ☐
Hotel accommodation close to city centre ☐
Airport transfers for £200 or less ☐
Total basket price £2000 or less ☐

# ACTIVITY 7

**Home > City Escapes > Filter page > Basket > Passengers Information**
To mark your birthday, you and your partner have decided to visit Amsterdam in November 2017 for a duration of 5 nights.

i. You wish to depart from London Heathrow using KLM Royal Dutch Airlines
ii. You want a twin standard room that includes bed and breakfast
iii. Upon arrival at Amsterdam, you require Airport transfer on a Luxury Eco Car to your hotel.

Your set budget for this trip is **£990** or less for both of you.
How can you accomplish this task?

**Stop where you are required to enter passengers' information.**

**Please tick the criteria you have met**

Depart from London Heathrow using KLM Royal Dutch Airlines

Twin standard room with bed and breakfast

A Luxury Eco Car to your hotel

Set budget for this trip is £990 or less

How do you accomplish this holiday booking with a set budget of **£2,300** or less?

**Please stop where you are required to enter passengers' information.**

**Please tick appropriately if you were able to meet the following criteria**

Flight from London, Gatwick airport

Hotel with safety deposit box

Flexible booking

Luggage allowance of 20kg for two adults

A set budget of £2,300 or less?

# ACTIVITY 8

**Home > Flight > Filter page > Basket > Passengers Information**

You are planning a 7nights summer holiday (May 2017) to visit Spain with your partner and two kids 5 and 8 years old that includes flight and hotel payment in one package.

- You wish to depart from London, Gatwick airport
- You want hotel accommodation with safety deposit box

**Please first complete the above tasks then include the following extras on your booking**

**Extras**

i. You need a flexible booking in case you would want to change your holiday.
ii. Outbound luggage allowance of 20kg for two adults

# ACTIVITY 9

**Home > Cruise > Mediterranean > Cruise search results > Cruise finder tool > Cruise search results > Passenger Information**

You wish to go on a Cruise to the Mediterranean on August 2017 for a duration of 10 to 14 nights and have chosen Southampton, England to be the port where the cruise ship will depart.

You **DO NOT** need a specific cruise line or cruise ship

i. You want to include airfare to and from Newcastle Intl. Airport, Newcastle
ii. You want a cruise that will visit at least 5 different ports
iii. You want a cabin with balcony and ocean view
iv. The total budget for the booking should be **£4,450**.

How can you accomplish these tasks?

**Stop where your booking meets all criteria including your set budget.**

**Please tick appropriately where you met the criteria**

Airfare
□
A cruise that will visit at least 5 ports
□
Cabin with balcony and ocean view
□
A set budget of £4,450 or less
□

# ACTIVITY 10

**Home > Extras > Contact Extras**

You have recently come back from a cruise holiday. During your time away you became unwell and ended up missing a shore excursion because you were stuck in your cabin for 3 days.

Luckily you took out Thomas Cook Silver insurance

i. See if you can claim the cost of this excursions which was £400
ii. Find out if you can claim any money for the 3 days of your holiday that you missed
iii. Find out the contact number/email for making a claim

**Please write you're your answers**

Can you claim the £400 cost of your

excursions:……
……

Can you claim any money for the 3 days holiday you have missed?

……..……
……

Contact number to make a claim

……
……..

Email address to make a claim

……
…….

**A5: Instrument for measuring participants testing experience**

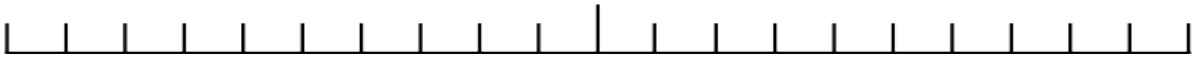
**A5.1: TLX Mental Workload questionnaire**

# TLX Mental Workload Questionnaire

Please tick the appropriate box on for each of the following questions that best described your experience with regards to the usability test experiment which you just complete

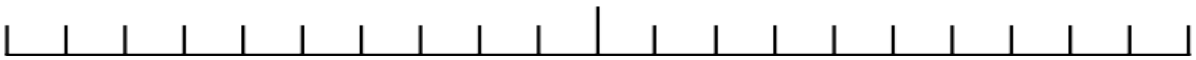
**Activity 1**
**Participant group:……**

1. Mental Demand: How mentally demanding was the task?



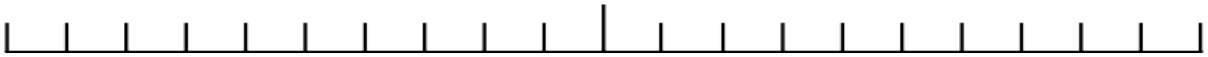
Very Low Very High

2. Temporal Demand: How hurried or rushed was the pace of the task?



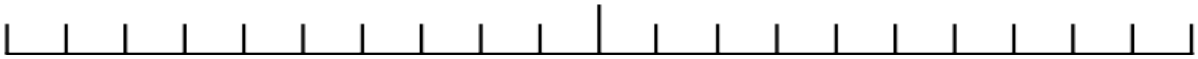
Very Low Very High

3. Performance: How successful were you in accomplishing what you were asked to do?



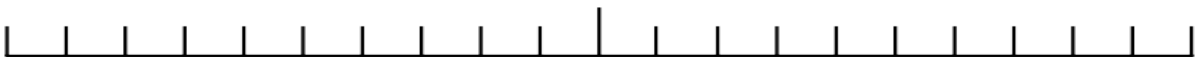
Very Low Very High

4. Effort: How hard did you have to work to accomplish your level of performance?



Very Low Very High

5. Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?



Very Low Very High

**A5.2: After Scenario Questionnaire for Both Classic and Silent condition**

# After Scenario Questionnaire

## Instructions

*Please circle your answer to the question using the provided 5-point scale (where 1 means strongly disagree and 5 means strongly agree).*

1. I feel satisfied with the number of successfully completed tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

2. I was able to concentrate during task performance

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

3. I was worried about talking too long over tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

4. The presence of the test facilitator made you feel uncomfortable

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

5. I persisted with tasks for longer than I would normally do so in real work use

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

6. I was concerned about giving up early on those tasks I found difficult.

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

Were any of these issues a factor in you giving up?

□ Bordon □ Time factor

□ Frustration □ Felt unachievable

7. I prefer to work in silence

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

**A5.3: After Scenario Questionnaire for Classic Only**

# After Scenario Questionnaire (TA)

## Instructions

*Please circle your answer to the question using the provided 5-point scale (where 1 means strongly disagree and 5 means strongly agree).*

1. Thinking aloud interfered with my performance during the tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

2. I found thinking-aloud during the tasks to be easy

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

3. The things I said during my think-aloud reflected all of my thoughts about the tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

4. I with held some information from my think-aloud

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

**A6: Demographic characteristics: Participant details**

| Participant Number | Gender | Age | Occupation | Duration of Internet use (years) | Frequency of Internet use | Frequency of travel website usage (3: always – 0: never) | Use of a travel website to look up different holiday destination in the past | Name of the website used | Used Thomas cook site before? | Usability testing experience? |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | Female | 18 | Undergraduate | 5 | 4 | 3 | Yes | Booking.com | No | No |
| P2 | Male | 25 | Undergraduate | 8 | 4 | 2 | Yes | Thompson.co.uk | No | No |
| P3 | Male | 18 | Undergraduate | 7 | 4 | 3 | Yes | Thompson.co.uk | No | No |
| P4 | Male | 19 | Undergraduate | 13 | 4 | 3 | Yes | secretescapes.com | No | No |
| P5 | Male | 21 | Undergraduate | 16 | 4 | 2 | Yes | travelzoo.co.uk | No | Yes |
| P6 | Male | 21 | PhD Student | 10 | 4 | 3 | Yes | Sky scanner | No | Yes |
| P7 | Male | 19 | PhD Student | 12 | 4 | 2 | Yes | Booking.com | No | No |
| P8 | Male | 23 | Undergraduate | 6 | 4 | 2 | Yes | Sky scanner | No | No |
| P9 | Male | 20 | Undergraduate | 11 | 4 | 3 | Yes | bestattravel.co.uk | No | Yes |
| P10 | Male | 18 | Undergraduate | 12 | 4 | 2 | Yes | PIA Airlines | No | No |
| P11 | Male | 30 | PhD Student | 20 | 4 | 3 | Yes | Norwegian.com | No | No |
| P12 | Female | 33 | PhD Student | 15 | 4 | 3 | Yes | Sky scanner | No | No |
| P13 | Female | 30 | PhD Student | 13 | 4 | 3 | Yes | Bookings.com | No | Yes |
| P14 | Male | 30 | PhD Student | 8 | 4 | 3 | Yes | Cheapair.com | No | Yes |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| P15 | Male | 32 | PhD Student | 17 | 4 | 3 | Yes | TUI UK | No | Yes |
| P16 | Male | 33 | PhD Student | 15 | 4 | 3 | Yes | Trainline | No | Yes |
| P17 | Male | 32 | PhD Student | 10 | 4 | 3 | Yes | Sky scanner | No | Yes |
| P18 | Male | 30 | PhD Student | 10 | 4 | 3 | Yes | Booking.com | No | Yes |
| P19 | Female | 26 | PhD Student | 9 | 4 | 2 | Yes | Sky scanner | No | No |
| P20 | Male | 27 | PhD Student | 18 | 4 | 3 | Yes | Booking.com | No | No |

Internet usage (4: daily; 3: Several times a week; 2: Several times a month; 1: once a month; 0: never)

**A7: After Scenario Question for both CTA and Silent**

| | CTA | SIL | Value |
|---|---|---|---|
| I feel satisfied with the number of successfully completed tasks | | | |
| I was able to concentrate during task performance | | | |
| I was worried about talking too long over tasks | | | |
| The presence of the test facilitator made you feel uncomfortable | | | |
| I persisted with tasks for longer than I would normally do so in real work use | | | |
| I was concerned about giving up early on those tasks I found difficult.<br>Bordon<br>Time factor<br>Frustration | | | |
| Felt unachievable | | | |
| I prefer to work in silence | | | |

**After Scenario Questionnaire for CTA**

| | CTA |
|---|---|
| Thinking aloud interfered with my performance during the tasks | |
| I found thinking-aloud during the tasks to be easy | |
| The things I said during my think-aloud reflected all of my thoughts about the tasks | |
| I with held some information from my think-aloud | |

**Appendix B: Materials from The Impact of Task-Type on Two Different Think-Aloud Protocols in Usability Testing Study**

**B1: Participant information sheet**

# Participant Information Sheet

**TITLE: A USABILITY EVALUATION OF THE NEXUS TRAVEL WEBSITE**
**(**https://www.nexus.org.uk/ **)**

***Before you decide to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully.***

**Purpose of the study**

To evaluate nexus transportation service website in terms of its user-friendliness.

**Background and aim of the study**

The study aims to conduct a usability evaluation on the above website. The study is concerned with testing the user friendliness of the above named website; it is not a test of your ability to use the site. During the test you will complete a number of everyday tasks, these might involve searching for information that exist on the website.

**Why have I been approached?**

You have been asked because you are a representative user for the website. This means you may use the website to find relevant transportation information.

**Do I have to take part?**

Participating in this study is entirely voluntary and that refusal or withdrawal will involve no penalty or loss, now or in the future.

**What will happen if I don't want to carry on with the study?**

You have the right to change your mind and withdraw from this study at any time without giving a reason and without incurring any penalties. If at any stage in the study you feel like to withdraw, just let the test facilitator know and ***all data collected up to the point of withdrawal will be immediately destroyed.***

**What will happen to me if I take part?**

The study will take place at the usability Laboratory of the University of Sunderland, and will involve you and the author. As you interact with the website the research will record a video of your computer screen and audio data. During the test you may be asked think-aloud and complete some questions about the tasks you have completed e.g. whether you found them difficult. The session will last about 40 minutes.

***Ethical review of the study***

The project has received ethical approval from the Research Ethics Committee (REC) of the University of Sunderland and it is being conducted in accordance with the University's Research Ethics Principles, Professional Codes of Practice and the law.

**What are the possible disadvantages and/or risks in taking part?**

There are no reasonably foreseeable discomforts, disadvantages, and risks associated with participating in this study.

**What are the possible benefits of taking part?**

Your participation will contribute to knowledge by providing usability practitioners with a deeper insight into the validity and reliability of their test data during usability testing.

**What if something goes wrong?**

If you are unhappy with the conduct of this study please contact myself Obruche Orugbo or my research supervisor Sharon McDonald, or the Chair of the University of Sunderland Research Ethics Group

Contact details are included below:

**Author**

Name: Obruche Orugbo

Email: bg35nr@research.sunderland.ac.uk

Phone 07438341102

**Research Supervisor**

Name: Sharon McDonald

Email: sharon.mcdonald@sunderland.ac.uk

**The Chair of Research Ethics Group**

Doctor John Fulton

Email: john.fulton@sunderland.ac.uk

**How will my information in this project be kept confidential?**

Information such as your gender, age, nationality, educational qualification, occupation and your use of the internet. Information will be kept in a secure locked cabinet or a password protected computer. Your responses e.g. transcripts of audio/video recordings or any other response data will be pseudo-anonymised using participant codes and kept separately from personal identifying information.

Data files will be kept for a period of two years and will be stored in a secure computer that is password protected.

The data may be looked at by staff authorised by the University of Sunderland for audit and quality assurance purposes

**What will happen to the results of this study?**

Results will be presented at conferences and written up in journals. Results are normally presented in terms of groups of individuals. The data collected during the course of this study might be used for additional or subsequent research if any individual data are presented; the data will be totally anonymous, without any means of identifying the individuals involved.

**Who is organising and funding the research?**

This research is organised by Obruche Orugbo, who is a full time research student at the University of Sunderland, Faculty of Computer Science, and School of Computer Science.

**Who has reviewed the study?**

The study has been reviewed and approved by the University of Sunderland Research Ethics Group.

**Further information and contact details**

Name: **Obruche Orugbo**

Email address: bg35nr@research.sunderland.ac.uk

Name of supervisor: **Sharon McDonald**

Email address: sharon.mcdonald@sunderland.ac.uk

Phone: 01915157385

**Dr. John Fulton** (Chair of the University of Sunderland Research Ethics Group)

Email address: john.fulton@sunderland.ac.uk

Phone: 01915152529

**B2: Participant Informed Consent Form**

**Thank you for taking time to read the information sheet!**

# Research Informed Consent Form

Study Title: **The Impact of Task Type on the Utility of Different Think-aloud Protocol in Usability Testing: a clarification between reliable utterances and reactivity**

*Please read the following statements and, if you agree, tick the corresponding box to confirm agreement:*

| | **Please initial box** |
|---|---|
| I confirm that I am over the age of 16 years. | ☐ |
| I have read and understood the information sheet for the above study and have had the opportunity to ask questions. | ☐ |
| I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason. | ☐ |
| I agree to take part in the above study. | ☐ |

| | **Please initial box** | |
|---|---|---|
| | Yes | No |
| I agree to the study being audio recorded. | ☐ | ☐ |
| I agree to the study being video recorded. | ☐ | ☐ |
| I agree to the use of anonymised quotes in publications. | ☐ | ☐ |

| Name of Participant | Date | Signature |
| --- | --- | --- |
| | | |
| Name of Author | Date | Signature |

If you would like a copy of this consent form to keep, please ask the author. If you have any complaints or concerns about this research, you can direct these, in writing, to the Research supervisor Dr. Sharon McDonald by email at: sharon.mcdonald@sunderland.ac.uk

**B3: User profile questionnaire**

# User Profile Questionnaire

*Please fill, tick and circle where appropriate*

**Demography**

i. Participant's Number:

ii. Gender: Male ☐ Female ☐

iii. Age:……...

iv. Nationality:.....

v. Highest educational qualification:

MSc ☐

BSc ☐

Senior Secondary ☐

vi. Occupation:.....

**Profile Questions**

1. How long have you been using the internet?.....

2. How often do you use the internet? Daily ☐
Several times a week ☐
Several times a month ☐
Once a month ☐

5. Have you used a public transport service website to plan a journey in the past? **Yes/No**

If yes, when was the last time?..... and what is the name of the transport company or website :.....

6. Have you used the Nexus website? **Yes/No**

7. Have you ever participated in usability testing? **Yes/No**

**B4: Tasks sets**

**Instruction: I would like you to think-aloud, tell me the things that you like, dislike or find confusing about the site as you complete each task**

**Activity 1**

Find and download the bus timetable for the Arriva bus X18 from Newcastle, Haymarket to Morpeth

**Answer Book**

Where you able to find and download the time table for Arrival bus X18?

- ☐ Yes
- ☐ No

**Activity 2**

Find out the cost of a student metro season ticket for one month, which covers all zones.

**Answer Book**

Please write the cost

…….

**Activity 3**

Find out if you can park your bicycle at Monument Metro station and the numbers of bicycle racks

**Answer Book**

Can you park your bicycle?

- ☐ Yes
- ☐ No

Please write down the number of bicycle racks

…….

**Activity 4**

When does the last ferry leaves North Shields on a Sunday?

**Answer Book**

Please write the time

…….

**Activity 5**

You want to apply for a pop card. Rather than apply online or download the application form you prefer to call a team

member for further enquiry. Find the phone number of their pop card team

**Answer Book**

Please write down the phone number

…….

**Activity 6**

You would like to use the **live travel map** to see the next available Go North East bus 700 from Sunderland University Travel Hub and download the timetable.

How would you accomplish this task?

**Answer Book**

Please write the time for the next available bus 700

…….

Were you able to download the timetable?

- ☐ Yes
- ☐ No

**Activity 7**

As a student at the University of Sunderland you usually use the metro to travel from **Newcastle** to **Sunderland**. You would like to get metro alerts via email if any disruptions occur between 7:00 and 15:00 for Mondays and Thursdays. How would you accomplish this task?

**Answer Book**

Were you able to create an email alert?

- ☐ Yes
- ☐ No

**Activity 8**

You wish to explore Tyne and Wear to visit places of attraction and have chosen to visit the "Angel of the North" with postcode **NE9 7UB**. Plan your journey from **Sunderland University Metro Station**.

What is the Go North East bus service number with the most convenient route to see the Angle of the North?

**Answer Book**

Please write down the Go North East bus number

…….

**Activity 9**

**Time and date for this activity is optional**

You and your step sister who is 15 years old both live at Jarrow close to the metro station. She will be attending South Tyneside College, South Shields. She has asked you to help her find out:

i. The most convenient route for her journey from home to college
ii. The cheapest return ticket

**Answer Book**

Write down the most convenient route?

…….

…….

…….

Please write down the name of the cheapest return ticket

…….

**Activity 10**

You live in **South Hylton** and wants to visit a course mate who lives at **Palmersville**. Since you are going on the metro, you want to buy a return ticket.

i. What are the names of the zones?
ii. What is the ticket's price?

**Answer Book**

Can you? Please write your answer

Please write down the zones

…….

Please write down the ticket's price

…….

**B5: Instrument for measuring participants testing experience**

**B5.1: TLX Mental Workload questionnaire for all conditions**

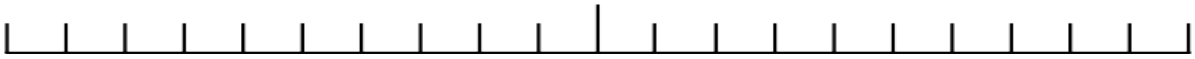
# TLX Mental Workload Questionnaire

Please tick the appropriate box on for each of the following questions that best described your experience with regards to the usability test experiment which you just complete

**Activity 1**

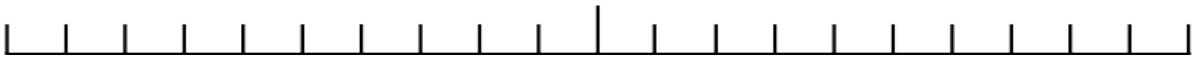
**Participant group**

1. Mental Demand: How mentally demanding was the task?



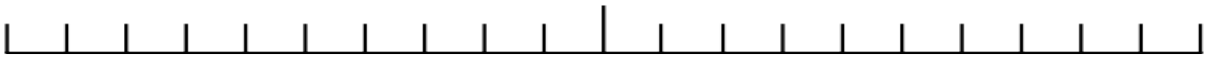
Very Low Very High

2. Temporal Demand: How hurried or rushed was the pace of the task?



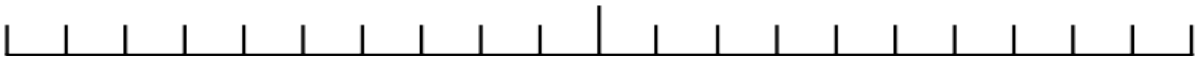
Very Low Very High

3. Performance: How successful were you in accomplishing what you were asked to do?



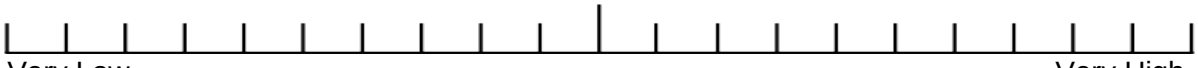
Very Low Very High

4. Effort: How hard did you have to work to accomplish your level of performance?



Very Low Very High

5. Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?



Very Low Very High

**B5.2: After Scenario Questionnaire for both Classic and Explicit condition**

# After Scenario Questionnaire

**Instructions**

*Please circle your answer to the question using the provided 5-point scale (where 1 means strongly disagree and 5 means strongly agree).*

1. I was able to concentrate during task performance

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

2. Thinking aloud interfered with my performance during the tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

3. I was worried about talking too long on those tasks I found difficult

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

4. The presence of the test facilitator made you feel uncomfortable

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

5. I persisted with tasks for longer than I would normally do so in real work use

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

6. The things I said during my think-aloud reflected all of my thoughts about the tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

7. I withheld some information from my think-aloud

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

8. I was concerned about giving up early on those tasks I found difficult.

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

Were any of these issues a factor in you giving up?

□ Burdon □ Time factor □ Frustration

□ Unachievable task □ others

9. I prefer to work in silence

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

10. I feel satisfied with the number of successfully completed tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

**B5.3: After Scenario Questionnaire for Classic Only**

# After Scenario Questionnaire (SIL)

**Instructions**

*Please circle your answer to the question using the provided 5-point scale (where 1 means strongly disagree and 5 means strongly agree).*

1. I was worried about talking too long on those tasks I found difficult

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

2. The presence of the test facilitator made you feel uncomfortable

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

3. I persisted with tasks for longer than I would normally do so in real work use

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

4. I was concerned about giving up early on those tasks I found difficult.

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

Were any of these issues a factor in you giving up?

□ Boredom □ Time factor □ Frustration

□ Unachievable task □ others

5. I feel satisfied with the number of successfully completed tasks

**STRONGLY DISAGREE** 1 2 3 4 5 **STRONGLY AGREE**

……..

……..

## B6: Demographic characteristics: Participant details

| Participant Number | Gender | Age | Occupation | Duration of Internet use (years) | Frequency of Internet use | Frequency of travel website usage (3: always – 0: never) | Use of a travel website to look up different holiday destination in the past | Name of the website used | Used Nexus travel website before? | Usability testing experience? |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | Male | 31 | Student | 10 | 4 | 4 | Yes | www.sl.se | Yes | Yes |
| P2 | Male | 30 | Student | 10 | 4 | 2 | Yes | Journey planner | No | No |
| P3 | Male | 31 | Student | 10 | 4 | 3 | Yes | SL.se (Swedish) | Yes | Yes |
| P4 | Male | 26 | Student | 10 | 4 | 4 | Yes | Nexus, Go North-East | Yes | No |
| P5 | Female | 28 | Student | 10 | 4 | 4 | Yes | Trainline | No | Yes |
| P6 | Female | 21 | Student | 10 | 4 | 3 | Yes | National Express | No | No |
| P7 | Male | 32 | Student | 8 | 4 | 2 | Yes | National Express | No | Yes |
| P8 | Male | 33 | Student | 10 | 4 | 2 | Yes | National Rail | No | No |
| P9 | Male | 27 | Student | 10 | 4 | 4 | Yes | Google Map | No | No |
| P10 | Female | 23 | Student | 10 | 4 | 4 | Yes | Google Map | Yes | No |
| P11 | Female | 21 | Student | 10 | 4 | 4 | Yes | Google Map | Yes | No |
| P12 | Male | 23 | Student | 10 | 4 | 3 | Yes | Sky scanner | No | Yes |
| P13 | Male | 24 | Student | 10 | 4 | 4 | Yes | National Express | No | No |
| P14 | Female | 32 | Student | 12 | 4 | 4 | Yes | Mega Bus | No | No |
| P15 | Male | 34 | Student | 10 | 4 | 3 | Yes | TUI UK | Yes | Yes |
| P16 | Male | 34 | Student | 10 | 4 | 3 | Yes | Go North-East | No | Yes |
| P17 | Male | 30 | Student | 10 | 4 | 3 | Yes | Mega Bus | No | Yes |
| P18 | Male | 37 | Student | 14 | 4 | 4 | Yes | Booking.com | Yes | No |
| P19 | Male | 31 | Student | 10 | 4 | 3 | Yes | National Express | No | No |
| P20 | Male | 34 | Student | 10 | 4 | 3 | Yes | Nexus | Yes | Yes |
| P21 | Male | 29 | Student | 10 | 4 | 3 | Yes | Trainline | Yes | No |
| P22 | Male | 34 | Student | 10 | 4 | 4 | Yes | National Express | Yes | Yes |
| P23 | Female | 26 | Student | 10 | 4 | 3 | Yes | National Rail | Yes | No |
| P24 | Male | 30 | Student | 10 | 4 | 4 | Yes | Google Map | No | No |
| P25 | Male | 30 | Student | 11 | 4 | 3 | Yes | Mega Bus | No | No |
| P26 | Male | 21 | Student | 10 | 4 | 3 | Yes | Nexus | Yes | No |
| P27 | Female | 40 | Student | 15 | 4 | 4 | Yes | Virgin Train | No | No |
| P28 | Female | 30 | Student | 10 | 4 | 3 | Yes | Nexus Website | Yes | Yes |
| P29 | Female | 39 | Student | 15 | 4 | 4 | Yes | Trainline | No | No |

| Participant Number | Gender | Age | Occupation | Duration of Internet use (years) | Frequency of Internet use | Frequency of travel website usage (3: always – 0: never) | Use of a travel website to look up different holiday destination in the past | Name of the website used | Used Nexus travel website before? | Usability testing experience? |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| P30 | Male | 29 | Student | 10 | 4 | 3 | Yes | Booking.com | No | No |
| P31 | Male | 30 | Student | 13 | 4 | 3 | Yes | Arrival | No | No |
| P32 | Male | 31 | Student | 10 | 4 | 2 | Yes | Go North East | No | No |
| P33 | Male | 28 | Student | 10 | 4 | 4 | Yes | National Express | No | No |
| P34 | Male | 25 | Student | 10 | 4 | 4 | Yes | Tyneside Ferry | Yes | No |
| P35 | Male | 34 | Student | 10 | 4 | 2 | Yes | Mega Bus | Yes | Yes |
| P36 | Male | 33 | Student | 10 | 4 | 3 | Yes | Thomas Cook | Yes | Yes |
| P37 | Male | 28 | Student | 10 | 4 | 4 | Yes | Trainline | No | Yes |
| P38 | Male | 28 | Student | 8 | 4 | 4 | Yes | Virgin Train | Yes | No |
| P39 | Female | 30 | Student | 10 | 4 | 3 | Yes | Traveline | Yes | No |
| P40 | Male | 27 | Student | 9 | 4 | 3 | Yes | Arriva | Yes | Yes |
| P41 | Male | 32 | Student | 10 | 4 | 2 | Yes | Google Map | No | No |
| P42 | Male | 26 | Student | 12 | 4 | 3 | Yes | British Airways | Yes | No |
| P43 | Male | 35 | Student | 10 | 4 | 3 | Yes | National Express | Yes | Yes |
| P44 | Female | 28 | Student | 13 | 4 | 3 | Yes | Grand Central | Yes | Yes |
| P45 | Male | 45 | Student | 10 | 4 | 3 | Yes | Arrival | Yes | No |
| P46 | Male | 35 | Student | 15 | 4 | 4 | Yes | Travel website | Yes | No |
| P47 | Male | 29 | Student | 15 | 4 | 4 | Yes | Metro & Google | Yes | No |
| P48 | Male | 28 | Student | 8 | 4 | 4 | Yes | Trainline | No | No |
| P49 | Male | 28 | Student | 10 | 4 | 4 | Yes | Google Map | Yes | Yes |
| P50 | Male | 35 | Student | 10 | 4 | 3 | Yes | Kayak | Yes | No |
| P51 | Male | 35 | Student | 10 | 4 | 4 | Yes | Arriva | No | No |
| P52 | Male | 30 | Student | 15 | 4 | 3 | Yes | National Express | Yes | No |
| P53 | Male | 32 | Student | 10 | 4 | 4 | Yes | National Express | No | Yes |
| P54 | Male | 20 | Student | 12 | 4 | 4 | Yes | Go North East | No | No |
| P55 | Male | 18 | Student | 10 | 4 | 3 | Yes | Transport for London | No | No |
| P56 | Female | 18 | Student | 8 | 4 | 3 | Yes | Trainline | No | No |
| P57 | Male | 30 | Student | 10 | 4 | 3 | Yes | Mega Bus | No | Yes |
| P58 | Female | 18 | Student | 10 | 4 | 4 | Yes | Nexus | Yes | No |
| P59 | Female | 18 | Student | 6 | 4 | 4 | Yes | Trainline | Yes | No |
| P60 | Female | 34 | Student | 18 | 4 | 4 | Yes | TUI | No | Yes |

Internet usage (4: daily; 3: Several times a week; 2: Several times a month; 1: once a month; 0: never)

## Appendix C: Information sheet

in this study.

**What are the possible benefits of taking part?**

Your participation will contribute to knowledge by providing usability practitioners with a deeper insight into the validity and reliability of their test data during usability testing.

**What if something goes wrong?**

If you are unhappy with the conduct of this study please contact me Obruche Orugbo or my research supervisor Sharon McDonald, or the Chair of the University of Sunderland Research Ethics Group Contact details are included below:

**Researcher**

Name: Obruche Orugbo

Email: bg35nr@research.sunderland.ac.uk

Phone 07438341102

**Research Supervisor**

Name: Sharon McDonald

Email: sharon.mcdonald@sunderland.ac.uk

**The Chair of Research Ethics Group**

Doctor John Fulton

Email: john.fulton@sunderland.ac.uk

**How will my information in this project be kept confidential?**

Information such as your ***gender, age, nationality, educational qualification, occupation and the use of think-aloud during a usability test***. Information will be kept in a secure locked cabinet or a password-protected computer. Your responses e.g. transcripts of audio/video recordings or any other response data will be pseudo-anonymised using participant codes and kept separately from personal identifying information. Data files will be kept for a period of two years and will be stored in a secure computer that is password protected. The data may be looked at by staff authorised by the University of Sunderland for audit and quality assurance purposes

**What will happen to the results of this study?**

Results will be presented at conferences and written up in journals. Results are normally presented in terms of groups of individuals. The data collected during the course of this study might be used for additional or subsequent research if any individual data are presented; the data will be totally anonymous, without any means of identifying the individuals involved.

**Who is organising and funding the research?**

This research is organised by Obruche Orugbo, who is a full-time research student at the University of

## Welcome to the research study!

We are interested in understanding the impact of think-aloud in a usability test, the purpose of the study is to investigate the impact of think-aloud in usability testing.

**Background and aim of the study**

The study will make use of an interview method to investigate factors that influence reactivity within a usability test by asking some related question about the think-aloud method, its use and implement during a usability test.

**Why have I been approached?**

You have been asked because you are a representative user and you have used the think-aloud method during a usability test session.

**Do I have to take part?**

Participating in this study is entirely voluntary and that refusal or withdrawal will involve no penalty or loss, now or in the future.

**What will happen if I don't want to carry on with the study?**

You have the right to change your mind and withdraw from this study at any time without giving a reason and without incurring any penalties. If at any stage in the study you feel like to withdraw, just let the test facilitator know and ***all data collected up to the point of withdrawal will be immediately destroyed.***

**What will happen to me if I take part?**

The interview session will be carried out on Microsoft Teams which will take place online, and will involve you and the researcher. As you answer the interview questions the researcher will record both the video and audio conversation. The session will last for about 30 minutes.

***Ethical review of the study***

The project has received ethical approval from the Research Ethics Committee (REC) of the University of Sunderland and it is being conducted in accordance with the University's Research Ethics Principles, Professional Codes of Practice and the law.

**What are the possible disadvantages and/or risks in taking part?**

There are no reasonably foreseeable discomforts, disadvantages, and risks associated with participating

Sunderland, Faculty of Computer Science, and School of Computer Science.

**Who has reviewed the study?**

The study has been reviewed and approved by the University of Sunderland Research Ethics Group.

**Further information and contact details**

Name: **Obruche Orugbo**

Email address: bg35nr@research.sunderland.ac.uk

Name of supervisor: **Sharon McDonald**

Email address: sharon.mcdonald@sunderland.ac.uk

Phone: 01915157385

**Dr. John Fulton** (Chair of the University of Sunderland Research Ethics Group)

Email address: john.fulton@sunderland.ac.uk

Phone: 01915152529

**Thank you for taking the time to read the information sheet!**

>>

Participant Signature

× SIGN HERE

clear

<< >>

Powered by Qualtrics

**C2: Participant Informed Consent Form**

Please answer this question.

## Research Informed Consent Form

Study Title: **To Investigate the Impact of Think-aloud in a usability test.**

**Please read the following statements and, if you agree, click all corresponding box to confirm an agreement:**

- ☐ I confirm that I am over the age of 16 years.
- ☐ I have read and understood the information sheet for the above study and have had the opportunity to ask questions.
- ☐ I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason.
- ☐ I agree to take part in the above study.

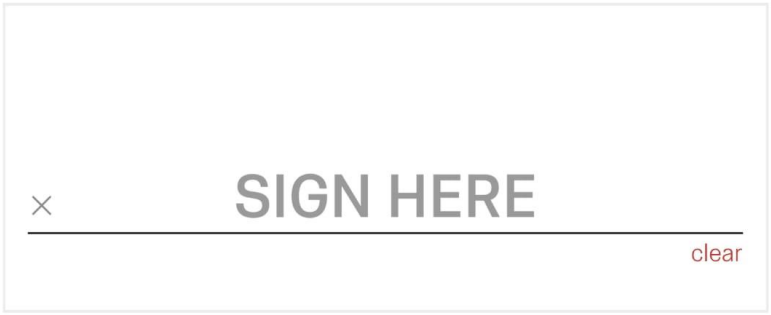
>>

Powered by Qualtrics

Click all corresponding box to agree to audio and video recording

- [ ] I agree with the study being audio recorded.
- [ ] I agree with the study being video recorded.
- [ ] I agree with the use of anonymised quotes in publications.

Participant Signature




<< >>

Powered by Qualtrics

## C3: Pre-screening Questionnaire

**PARTICIPANT PRE-SCREENING QUESTIONNAIRE**

**Please fill and tick where appropriate**

Please enter your full-name

Whats your highest level of educational qualification ?

- ☐ Research Degree PhD
- ☐ Masters Degree
- ☐ Bachelor's Degree

How long have you worked in User Research/Usability

Are you involved in Usability Test

- ○ Yes
- ○ No

How long have you been involved in Usability Test

Do you use Think-aloud

- ○ Yes
- ○ No

Thank you for taking your time to complete this questionnaire

>>

## C4: Interview Questions

### Think-aloud

1. Do you see think-loud as an essential part of usability testing?
2. What think-aloud method do you use and why?
3. Do you establish a rapport with participant, and do you think that is important?
4. When using the think-aloud method, do you ask users to practice thinking aloud? How and why?

### Intervention - Probe, Prompt

5. During think-aloud what would you normally expect people to talk about and what do you hope to find out?
6. Do you give participants specific type of think-aloud instructions? What kind of instructions?
7. Do you prompt participants during think-aloud sessions? How and when?
8. Let's talk about your experience when using the think-aloud for usability evaluation, when do you use it, are you using it for formative or summative evaluation or both?
9. Do you use it to identify and fix UX problems or to improve the quality of the User Experience?

### Tasks

10. What type of tasks do you ask people to work on during think-aloud sessions?
11. How do give tasks to participants?
12. Do you think asking participants to think-aloud makes them want to explore more than they will normally do?
13. How do you help a struggling participant move on or skip a difficult task?
14. What do you think are the limitations of using think-aloud?

**Reactivity**

15. Do you think asking people to think-aloud changes what they do? How?
16. When people think-aloud what kind of interactions do you have with them?
17. Do you consider observing participants important?
18. Do you stay in the same room with a participant during a usability test?
19. What do you do when people fall silent when using the think-aloud method?

**Data analysis**

20. What kind of performance metrics do you normally measure?
21. How do you analyse the data obtained?
22. What kind of information do you look for when analysing the data obtained from a think-aloud session?
23. How easy is it to analyse data obtained from a think-aloud session?
24. Is there a particular reason why you do it that way?

### C5: Demographic characteristics: Participant Profile

| No | Qualification | Work Experience | Usability Test | Think-aloud Usage | Work role | Location | Organisations |
|---|---|---|---|---|---|---|---|
| **1** | Ph.D. | 20 years | Yes | Yes | UX lead | UK | UX Consulting |
| **2** | Degree (BSc) | 5.5 years | Yes | Yes | UX Researcher | UK | Marketing |
| **3** | Ph.D. | 8 years | Yes | Yes | UX Researcher | UK | UX Consulting |
| **4** | Degree (BSc) | 7 years | Yes | Yes | UX Researcher | UK | Software Development |
| **5** | Degree (BSc) | 5 years | Yes | Yes | UX Researcher | UK | Software Development |
| **6** | Ph.D. | 6 years | Yes | Yes | UX Researcher | UK | Banking |
| **7** | Degree (BSc) | 5+ years | Yes | Yes | UX Researcher | UK | Software Development |
| **8** | Degree (BSc) | 5+ years | Yes | Yes | UX Researcher | America | Gaming |
| **9** | Degree (BSc) | 5 years | Yes | Yes | UX Researcher | Europe | Banking |
| **10** | Degree (BSc) | 20 years | Yes | Yes | UX Researcher | UK | Gaming |
| **11** | Masters (MSc) | 5 years | Yes | Yes | UX Researcher | Indian | Software Development |
| **12** | Degree (BSc) | 10+ years | Yes | Yes | UX Researcher | UK | Software Development |
| **13** | Degree (BSc) | 17 years | Yes | Yes | UX Manager | UK | Software Development |
| **14** | Degree (BSc) | 6-7 years | Yes | Yes | UX Researcher | UK | Marketing |
| **15** | Degree (BSc) | 18 years | Yes | Yes | UX Researcher | UK | Software Development |
| **16** | Degree (BSc) | 8 years | Yes | Yes | UX Researcher | UK | Healthcare |
| **17** | Masters (MSc) | 8 years | Yes | Yes | UX Researcher | UK | Software Development |
| **18** | Masters (MSc) | 5 years | Yes | Yes | UX Researcher | Europe | Supply Chain |
| **19** | Degree (BSc) | 7 years | Yes | Yes | UX Researcher | Europe | Healthcare |
| **20** | Degree (BSc) | 11 years | Yes | Yes | UX Researcher | America | Housing |
| **21** | Masters (MSc) | 5 years | Yes | Yes | UX Researcher | UK | Telecommunication |
| **22** | Degree (BSc) | 10+ years | Yes | Yes | UX Manager | UK | Telecommunication |