

**Wanda: A Tool to Support Remote Usability
Evaluations Using Think Aloud and
Question-Asking Protocol**

Daniel Saldarriaga López



Master of Science
Informatics
School of Informatics
University of Edinburgh
2022

Abstract

The rapid expansion of the internet has raised the awareness that development teams must provide ever better digital experiences. As a result, evaluating the usability of their digital products has become a requirement for many companies across industries. However, the available tools and methods for conducting usability evaluations are frequently costly, and many businesses cannot afford them. Furthermore, since the COVID-19 pandemic, the demand for remote usability evaluations has increased. Therefore, enterprises and usability researchers require a tool to conduct remote usability studies at an affordable price. In this project, I describe how I designed, developed and evaluated Wanda, an open-source tool that supports remote usability evaluation studies using Think Aloud [57] and Question-Asking Protocol [54]. I present the methodologies used for gathering the requirements for the tool, including a literature review, a benchmark on existing tools and interviews with HCI experts from The University of Edinburgh. In addition, I explain how I developed and evaluated two iterations of Wanda to understand its potential impact. In the end, after reaching a minimum viable product for Wanda and evaluating it with more than 28 potential users, I concluded that Wanda is a tool with great potential to support remote usability studies.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee. Ethics application numbers: 6718 and 349375. Date when approval was obtained: 2022-05-31 and 2022-07-25. The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Daniel Saldarriaga López)

Acknowledgements

I would like to thank the HCI experts and the participants that helped me in my studies. I would also like to thank my supervisor, Cristina Alexandru, for her assertive suggestions and great support during the project. Lastly, I would like to thank my family for their moral support.

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research goals	2
1.3	Report structure	3
2	Methodology	4
2.1	Background and related work	4
2.2	Requirement gathering and design	4
2.3	First iteration of development	5
2.4	Formative evaluation	6
2.5	Second iteration of development	7
2.6	Summative evaluation	7
3	Background and related work	9
3.1	Usability	9
3.2	Usability evaluation	10
3.2.1	Think Aloud [57]	11
3.2.2	Question-Asking Protocol [54]	13
3.2.3	System Usability Scale	14
3.2.4	Questionnaires	15
3.3	Related Work	15
4	Requirement gathering and design	17
4.1	Aims and Objectives	17
4.2	Protocol	17
4.3	Data analysis	18
4.4	Results	19
4.5	Discussion	20

4.5.1	Real life challenges of Think Aloud and Question-Asking Protocol	20
4.5.2	Advantages and disadvantages of existing tools	20
4.5.3	Design of Wanda	22
5	First iteration of development	23
5.1	Development stack	23
5.2	Back-end	24
5.3	Front-end	25
5.3.1	Expert Interface	25
5.3.2	Participant Interface	27
6	Formative evaluation	28
6.1	Aims and Objectives	28
6.2	Protocol	28
6.3	Data analysis	29
6.4	Results	29
6.5	Discussion	30
7	Second iteration of development	31
7.1	Participant Interface	31
7.2	Expert Interface	32
7.3	Discussion	33
8	Summative evaluation	34
8.1	Aims and Objectives	34
8.2	Protocol	34
8.3	Data analysis	35
8.4	Results	35
8.5	Discussion	36
9	Discussion, future work and conclusions	37
9.1	Discussion	37
9.1.1	Limitations	38
9.2	Future work	39
9.3	Conclusions	39
	Bibliography	41

A	Background and related work	52
A.1	Usability inspection methods	52
A.2	Usability evaluation tools	53
A.2.1	Evaluated usability testing tools	53
A.2.2	Additional usability evaluation tools	54
B	Requirement gathering and design	56
B.1	Initial sketch of Wanda	56
B.2	Participant recruitment	56
B.3	Data collection method	57
B.4	Materials	57
B.4.1	List of questions	57
B.5	Procedure	59
B.6	Data analysis	59
B.7	Results	60
B.7.1	Think Aloud and Question-Asking Protocol methodology	60
B.7.2	Feedback about tool features	67
B.7.3	User Stories	68
B.8	User Stories	69
B.9	Prioritized requirements	71
C	First iteration of development	74
C.1	Back-end	74
C.1.1	E-mails	74
C.2	Authentication	74
C.2.1	Table definition	75
C.2.2	Database ER Diagram	76
C.3	Front End	76
C.3.1	Landing and Authentication	77
C.3.2	Participants' interface	77
C.3.3	Experts' interface	77
D	Formative evaluation	88
D.1	Participant recruitment	88
D.2	Data collection method	88
D.3	Materials	89

D.3.1	Scripts	89
D.4	Procedure	95
D.5	Data analysis	95
D.6	Results	96
D.6.1	Participant interface	96
D.6.2	Expert interface	99
D.7	Updated list of requirements	102
E	Second iteration of development	104
E.1	Participants' interface	104
E.2	Experts' interface	104
F	Summative evaluation	113
F.1	Participant recruitment	113
F.2	Data collection method	113
F.3	Materials	114
F.3.1	Scripts	114
F.4	Procedure	126
F.5	Data analysis	126
F.6	Results	127
F.6.1	Participant interface	127
F.6.2	Expert interface	131
F.7	System Usability scale results	134
F.7.1	Participants	134
F.7.2	Experts	134
F.8	Future feature requirements for Wanda	134
G	Participants' information sheet	135
G.1	Requirement gathering	135
G.2	Formative evaluation	140
G.2.1	Experts	140
G.2.2	Participants	145
G.3	Summative evaluation	150
G.3.1	Experts	150
G.3.2	Participants	155

H	Participants' consent form	160
H.1	Requirement gathering	160
H.2	Formative evaluation	162
H.3	Summative evaluation	164

Chapter 1

Introduction

1.1 Motivation

Since the early 90s, the Internet has been growing at an accelerated pace, reaching points where we, as a society, can no longer run businesses, check our finances or even maintain our personal relationships without having constant access to the Internet [36]. That is why, now more than ever, the software we use must provide outstanding experiences [74]. Therefore we, as software developers, must be willing to begin a continuous improvement process on the digital experiences we design, develop and deploy.

To assess if users are satisfied with the digital products they use, we must take into account the concept of *usability*, which is defined by the ISO 9241-11 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*” [1].

Traditionally, usability evaluation has been performed using a wide range of methods that can be executed in different phases of the development of a system [65]. However, many businesses only use them in the late stages of development, owing to their high cost [24]. Several academics have investigated the impact of late-stage usability evaluation and have mostly agreed that late changes to the user interface are typically costly and time-consuming [21] [94] [13]. In many cases, usability evaluations are ignored altogether, leading to high costs due to development reprocesses in future iterations of the products [27]. As a result, it has become critical to incorporate usability evaluations as early as possible in the development stages of a project.

Usability evaluations have been typically implemented in in-person setups, where a usability expert can analyse how a user interacts with a digital product. However,

the challenges posed by the COVID-19 pandemic increased the necessity to evaluate usability in a remote setting [20], which led to a rise in online tools that provide features for conducting remote usability evaluations. Such tools usually allow an expert to set up an app or website they want to evaluate, and allows them to recruit participants for assessing how their experiences can be improved.

Therefore, this project aims to explain the design, development and evaluation of Wanda, a free, open-source tool that will allow experts to conduct remote usability evaluations of websites at any post-design stage using the Think Aloud [57] and Question-Asking Protocol [54] methods. Two methods that have been widely studied and whose contributions are well documented.

1.2 Research goals

The main objective of this project is to develop a tool that allows usability experts to design, conduct and analyse usability studies on websites, specifically using the Think Aloud [57] and Question-Asking Protocol [54] methods.

The goal of the project can be broken down into the following research questions:

1. What are the real-life challenges of Think Aloud [57] and Question-Asking Protocol [54]?
2. What are the advantages and disadvantages of current online usability evaluation tools using either Think Aloud [57] or Question-Asking Protocol [54]?
3. How can we design a tool that helps experts in conducting Think Aloud [57] and Question-Asking Protocol [54] evaluations?
4. How can we develop a tool that helps experts in conducting Think Aloud [57] and Question-Asking Protocol [54] evaluations?
5. What are the perceptions of potential users about:
 - (a) the usability of the developed tool?
 - (b) its potential impact of the developed tool in terms of:
 - i. easing the process of analysing evaluation data?
 - ii. encouraging participants to provide more feedback on the evaluated system?

1.3 Report structure

This dissertation is divided into nine chapters, structured as follows:

In chapter two, I show all the methodologies used throughout the project to design, develop and evaluate Wanda. I mention the technologies used in the project and the different strategies I followed for gathering feedback from the potential users of Wanda.

In chapter three, I explore the recent and historical literature to analyse how the definitions of usability have evolved and which methods have been proposed for usability evaluations. I also explore different tools that are already existing in the market, showing their characteristics and main functionalities.

In chapter four, I show the requirement gathering procedure and its results. I explain all the feedback given to me by five usability experts I interviewed to know their pain points when using the Think Aloud[57] and Question-Asking Protocol [54] methods.

In chapter five, I explain how I transformed the list of requirements obtained from the requirement gathering into the first iteration of Wanda, and I show the different technologies and technical decisions I made throughout the development.

In chapter six, I discuss the procedure and results of a formative evaluation that I did with three usability experts and ten potential participants of a usability study. I explain how they used the first iteration of Wanda to design, conduct and analyse an example of an evaluation study, and I expose which problems they found and what features they liked. I also show the score the experts and participants gave Wanda in terms of the System Usability Scale [16].

In chapter seven, I explain how I built the second iteration of Wanda, considering features that were missing from the first iteration and features that required adjustments after learning the results from the formative evaluation.

In chapter eight, I discuss the procedure and results of a summative evaluation that I conducted with five usability experts and fourteen potential participants divided into three focus groups. I show how Wanda behaves on a broader test with more people. I also explain which features were liked by the participants and which issues remain present in Wanda. I also show the score the experts and participants gave Wanda regarding the System Usability Scale. [16].

Finally, chapter nine presents my conclusions and thoughts regarding the project's results. I analyse the different results and present a summary of the answers to the research questions. I also discuss the limitations of this work and the potential future work that may come in subsequent projects further developing Wanda.

Chapter 2

Methodology

This chapter presents the methodology used in every phase of the project. It details the methods, plans and tools used in each section and which research questions are aimed in each section.

2.1 Background and related work

The goal of the background and related work phase was to lay down the theoretical concepts required for understanding how the different usability evaluation methods are defined. Additionally, this phase is intended to provide insights for answering research questions RQ1 and RQ2.

First, a literature review was conducted using search engines such as Google Scholar [39] and DiscoverEd [109]. On both search engines, I used mainly the following search keywords: *usability*, *usability evaluation*, *usability testing*, *“think aloud” usability*, *question asking protocol*, *usability testing tool*, *system usability scale*, amongst others.

Second, I also studied which tools were available on the market for usability research. I found more than fifteen tools that can potentially be used in different stages of the development, but I filtered them to the ones closer to the scope of this project and ended up using five of them to test how they work and which are their main features.

I discuss the background and related work in chapter 3.

2.2 Requirement gathering and design

The goal of the requirement gathering phase was to answer the research questions RQ1, RQ2 and RQ3.

Before knowing formal requirements from HCI experts, and based solely on the literature and market benchmarks explained in chapter 3, I created an initial draft of how I envisioned the tool. Using Figma [33], I produced a very early hand-drawn draft that was used to create a mental map of what Wanda could potentially look like, which can be seen in the appendix B.1.

With the initial draft and the knowledge from the literature and market, I designed a semi-structured interview [73] that was used to ask HCI experts their opinions on which functionalities the tool should have. I decided to use semi-structured interviews because each expert may have different experiences and opinions on Wanda's requirements, so the semi-structured interviews allowed me to ask further questions on a specific topic if required.

I conducted five 30-minute online one-to-one sessions using Microsoft Teams [108] with lecturers of The University of Edinburgh with vast experience in HCI. During the interviews, I asked them about their previous experience with both Think Aloud [57] and Question-Asking protocol [54], their pain points with their current way of executing both methods with users and their opinions on potential features for the tool.

The results from the interviews produced more than two hours of verbal feedback. I used Nvivo [87] for processing the transcripts because it provides features for quickly categorising the participants' ideas. I also chose to use Thematic Analysis [19] because it allowed me to cluster the different comments given by the participants and the experts on common topics, which I could then use to identify the most important features or changes I needed to consider for the second iteration of development. The combination of Nvivo and Thematic Analysis allowed me to obtain sufficient detail of the experts' opinions for forming a vision of their real struggles with the aforementioned methods.

At the end of this initial study, I built a detailed list of 23 requirements in the form of User Stories and prioritised them, considering how long they would take and how often the experts requested them.

More detail about the requirement gathering phase can be found in chapter 4.

2.3 First iteration of development

Based on the requirements gathered in the previous phase, I discuss in this step the details of Wanda's first iteration of development in this section. The information obtained in this project phase helps answer the research question RQ4.

Wanda is developed using Next.js [77], a React.js [98] framework that I chose

because I had previous experience with it and because it allows the development of the front-end and the back-end in the same repository, reducing the amount of code that needs to be maintained.

The first iteration of development focused on developing the platform's back-end and infrastructure and the interfaces for the experts and participants. At the end of this project stage, experts were able to create, conduct and analyse evaluation studies, and participants could connect to a session and use Think Aloud [57] for evaluating a website. Due to time constraints, the Question-Asking Protocol [58] method was left for the second iteration of development.

More details about the first iteration of the development and all the technical decisions made can be found in chapter 5.

2.4 Formative evaluation

This phase aimed to understand users' opinions on the features developed in the first iteration. A formative evaluation was conducted to understand how users interact with Wanda and how it can be improved, and also to help me build an answer to RQ4.

On the participant side, 10 MSc students from The University of Edinburgh were invited to one-to-one meetings via Microsoft Teams [108]. They were asked to use Wanda and execute four tasks using Think Aloud [57] for evaluating an e-commerce platform. To stimulate participants to give as much feedback as possible, I wanted them to see a new e-commerce website that they had not seen before. For that reason, I decided not to use common e-commerce sites such as Amazon [4] or e-bay [29]. Instead, I developed a mock e-commerce platform using Vercel's E-Commerce template [121] and filled it with mock products.

After the Think Aloud [58] session, I asked the students about their experience using Wanda during a semi-structured interview. I also asked them to fill out the System Usability Scale [16] questionnaire, which I chose because it provides a quick way of assessing the system's usability. It is also helpful because I can compare the SUS score in each iteration to see if the system is improving.

On the expert side, three HCI experts were invited to join one-to-one meetings via Microsoft Teams [108] in which they were asked to evaluate Wanda using Think Aloud[57]. During the Think Aloud session, they had to execute seven tasks in Wanda while externalising all their thoughts. After the session, I asked them questions during a semi-structured interview and to fill out the System Usability Scale [16].

After the sessions, I proceeded to extract the transcripts from the recordings and analyse all the data using Thematic Analysis [19] in NVivo [87], similar to how it was done in the requirement gathering phase.

Further detail about the formative evaluation can be found in chapter 6.

2.5 Second iteration of development

The goal of this phase was to present the most significant changes made to Wanda after the feedback obtained in the formative evaluation. This project phase also aims to answer the research question RQ4.

After the formative evaluation, the list of requirements was reviewed. Some features were added, and some were changed to adjust to the users' needs. Therefore, this project phase explains the most significant changes made to Wanda and the thought process behind developing a separate open-source package for visualising the System Usability Scale [16], which was not planned at the beginning but can also be a valuable outcome of this project.

More detail for the second iteration of the development can be found in chapter 7.

2.6 Summative evaluation

This phase aimed to understand what users believe about Wanda's final iteration of development, providing a summative view of the whole development process. In addition, this evaluation aimed to obtain the information required for answering the research question RQ5.

For evaluating the expert interface summatively, a similar method to the formative evaluation was implemented. In total, five experts were invited to one-to-one sessions via Microsoft Teams [108] where they had to execute eight tasks for designing, conducting and analysing the results of an evaluation study. After they finished the tasks, I asked them to fill out a questionnaire that contained the System Usability Scale [16] and some questions about Wanda's potential impact.

Finally, I targeted more students to evaluate the participants' interface, so it was unfeasible due to time constraints to conduct one-to-one sessions. That is why I divided the fourteen participants into three group studies conducted on Microsoft Teams [108], in which they were able to execute an evaluation of the same e-commerce platform

mentioned before. For keeping a common baseline, in this iteration I used the same e-commerce platform I had developed for the formative evaluation.

After the participants executed (in silence) each of the four tasks I gave them, I asked them questions about the features and motivated them to discuss with the other participants in the meeting. After the session ended, I sent the participants a questionnaire that also included questions about the System Usability Scale [16] and some questions about Wanda's potential impact.

Further insights can be found for the summative evaluation in chapter 8.

Chapter 3

Background and related work

In this chapter, I will review the academic literature relevant to this project. I will first start by explaining the most relevant definitions of usability and how it has evolved throughout the years. I will then expose the most relevant methods for evaluating the usability of a system and how they differ from each other. Finally, I will give more detail about the methods relevant to this project: Think Aloud [57], Question-Asking Protocol [54] and System Usability Scale [16].

3.1 Usability

Since we started designing and using software with graphic interfaces, we have been increasingly interested in providing better digital experiences. Since the 1970's we have been using the term *usability* as a synonym for *user-friendliness*, which is a rather broad term but exemplifies what users expect from a system: they want to be able to use it without any problems.

The earliest definitions of usability in the 1970s and 1980s were diverse and contained many different approaches [82]. The most relevant definitions at the time included defining *usability* as a factor of three different variables: the system's ergonomics, the mental effort required by the user to use the system, or how easy the system is [82].

Jakob Nielsen, widely regarded as the father of usability, provided in 1994 a more succinct definition of usability focusing on five components that every usable system should include [78]. First, the system should be *learnable*, meaning that a user should be able to learn how to use it effortlessly. Second, the system should be *efficient*, so a user should be highly productive when using it. Third, the system should be *memorable*, meaning that a user should be able to recall how to use the system even after a long

time of not using it. Fourth, a system should be *robust* enough to keep the error rate as close to zero as possible, and when an error does happen, it should allow users to recover from it. Finally, a usable system should be *satisfactory* to use, so users should feel subjectively satisfied when using the system.

Nowadays, we have accepted and standardized definitions of usability. The ISO 9241-11 defines usability in Human-Computer Interaction 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*” [1]. That definition emphasizes the effectiveness and satisfaction of the usage, but it is not far away from the one proposed by Nielsen more than 30 years ago.

3.2 Usability evaluation

The fact that we can define and understand usability does not necessarily mean that we can easily assess whether or not a system is usable. For doing so, we need to rely on well-defined procedures called *Usability Evaluation Methods* (UEMs), which allow us to assess the usability of a product in a systematic and repeatable way [45].

Usability evaluation methods have evolved throughout the years, and new methods have appeared, but they can be classified into two different categories first introduced by Nielsen in 1994 [79]. As cited by Paz et al. [91], Nielsen describes that the usability of an interface can be assessed using *inspection methods* or *testing methods* [78].

As explained by Nielsen, usability inspection [79] is the name given to the set of usability evaluation methods based on an expert directly analysing the system. Several methods can be found in this category: heuristic evaluations [79], cognitive walkthroughs [124], feature inspections [79] and pluralistic walkthroughs [12], amongst others. In 2007, Hollingsed et al. [47] reviewed the original inspection methods proposed by Nielsen and concluded that heuristic evaluation and cognitive walkthroughs were the most widely-used inspection methods. They thoroughly reviewed different literature sources and showed that inspection methods are not always reliable because even if usability experts perform them, not all the issues may be found, and it is still prone to human errors. Some brief descriptions of inspection methods can be found in the appendix A.1.

Unlike usability inspection, usability testing is not performed by an expert, but by a potential user of the evaluated system [91]. As explained by Riihiahho [99], usability testing methods usually involve a participant executing some previously defined tasks

within a system. Those tasks are defined by an expert who usually guides the participant through the evaluation.

Usability tests are based on four phases: first, an expert designs and prepares all the tasks the participant must execute. Next, they recruit participants and conduct the evaluations. Subsequently, the experts analyse the results, and finally, they communicate them [91].

Usability testing methods are broad, and there are several variations of the different methods. In the coming subsections, I will provide a more detailed review of the more relevant methods for this project.

3.2.1 Think Aloud [57]

Think Aloud is a usability evaluation method that was first proposed by Clayton Lewis in 1982 [57] based on the work by Ericsson and Simon in the field of psychology [32]. It is a simple yet powerful method in which users follow a series of tasks within a website or application. While they execute their tasks, they are supposed to speak or exteriorise what they think about the system while using it.

3.2.1.1 Advantages and disadvantages of Think Aloud

I will discuss in this section the different advantages and disadvantages of Think Aloud according to different academic sources. The information gathered from the literature aids in building a response to the research question RQ1.

One of the main benefits of [57] is that it allows experts to identify the users' thoughts as they happen, thus eliminating the issue of short-term memory loss [123]. Instead of Think Aloud [57], an expert could ask a participant to use a particular system and ask some questions after the participant has finished using the system. In such cases, the participants may not recall all the details of what they did and may not remember specific things they disliked, which leads to incomplete and inaccurate evaluations. Since the idea of Think Aloud [57] is that participants express their thoughts as they happen, then the experts can collect all the required feedback from the participants.

The Think Aloud [57] method can also reveal data that would not be available in other methods. When participants use a system and simultaneously express their thoughts, researchers can directly identify why a participant is using a particular feature in an unexpected way [22].

However, Think Aloud [57] also comes with some drawbacks. First, it is a method

that can not be used to assess the *efficiency* of a system because efficiency is typically linked to how fast a user can execute a particular task. Since during Think Aloud [57], participants need to speak while they execute their tasks, the time they take to finish them is not the accurate time they would have taken if they were not speaking [119].

Additionally, some participants may find that Think Aloud [57] is unnatural or uncomfortable, which may alter the way they perceive the interface and may also bias their overall opinion towards the experience of the system they are using [3].

Some experts also argue that the Think Aloud [57] method interferes with the participants' thought process, refraining them to unleash their full potential to execute the task. In other words, since participants need to speak aloud, they might find a task more difficult than it would be if they were not speaking, because their brains need to process both things at the same time [128] [46].

It has also been argued that Think Aloud [57] leads experts to bias the evaluation towards already-known issues instead of addressing new ones [85]. For example, if a participant struggles to complete a task, the evaluator might be tempted to hint at the solution. Such situations happen, especially when many participants fail on the same task. This practice prevents the participant from discovering the solution to their issue and thus also prevents the expert from identifying potential new failure methods [89].

Many authors also express concerns about the difficulty of analysing Think Aloud [57] data [106]. All the information that comes out of a Think Aloud [57] session is verbal and thus is non-structured. Experts need to transcribe the audio recordings of the participants' thoughts, and then they usually categorise the verbalisations into different topics using Thematic Analysis [19]. Therefore, Think Aloud [57] evaluations may take significantly more time than other methods with structured data, such as surveys [81].

Despite its drawbacks and the fact that it is a relatively old method, Think Aloud [57] is still widely used. Examples can be found in various industries and areas of knowledge. For instance, it has been implemented to assess the usability of healthcare software [125] [110], learning systems [18] [50] [107], financial applications [31] [127], amongst others. The fact that Think Aloud [57] is easy to use and its efficiency are widely documented makes it a good choice for companies that need to evaluate their systems. However, as further explained in section 3.3, there are not many widely-used Think Aloud [57] online tools that help experts conduct their remote studies using this method. That means that experts are usually restricted to finding participants within their physical locations [37], which is becoming more complex, considering post-pandemic societal behaviours [20].

3.2.2 Question-Asking Protocol [54]

The Question-Asking Protocol is a method first proposed by Takashi Kato in 1986 [54]. It is similar to Think Aloud [57] but adds additional layers of formality by allowing participants to ask explicit questions depending on the system's status while they execute the tasks.

In a typical Question-Asking Protocol session, an expert would also give tasks to the participant. While the participant is executing the task, they are encouraged to ask questions about the system's status. Asking questions allows experts to identify the thought process the participant is following [54].

According to the method proposed by Kato [54], the queries from the participants can be divided into two categories, which allows for rapid classification of the users' thoughts, which in turn allows for faster processing of the evaluation data. The classification is presented in figure 3.1.

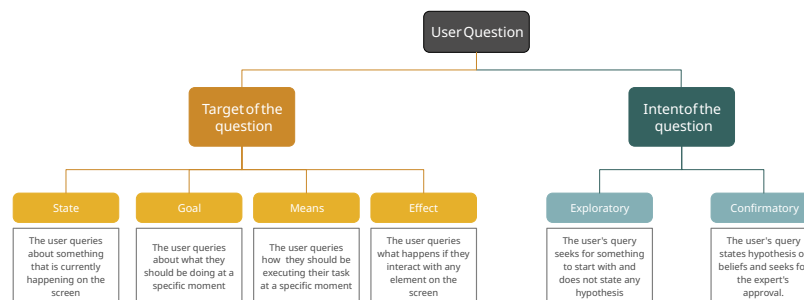


Figure 3.1: Question-Asking Protocol classification. Self built, first presented on [102], based on [54]

Rodrigues et al. [100] recently used the Question-Asking Protocol method and adapted it for analysing how blind people adapt to tutorials on smartphone apps. However, they did not use the method exactly proposed by Kato [54]. Instead, they followed guidelines from Grossman et al. [43] where they propose a modification to the original method to introduce the *question-suggestion* protocol, which gives more freedom to the expert to guide the participant towards the desired outcomes. It also allows the expert to ask questions to the participant, thus generating a two-way communication that can be more effective than the original method.

3.2.3 System Usability Scale

As stated before, both the Think Aloud [57] and Question-Asking Protocol [54] methods suffer from criticism due to the tedious process of analysing the participants' data. One of the most widely adopted methods for obtaining fast usability insights is the System Usability Scale (SUS), a method proposed by John Brooke in 1996 [16]. As Brooke defines it, the SUS is a "quick and dirty" method for obtaining a low-cost assessment of the usability of a system.

The System Usability Scale is a questionnaire based on a Likert scale [61] and contains ten questions designed to assess a system's usability. All questions have five possible choices, ranging from *strongly disagree* to *strongly agree* and are intended to be answered by a participant after interacting with a system.

The SUS was designed to have five questions with common *strongly agree* and five with common *strongly disagree* answers [16]. It is important to note that, according to Brooke [16], when the questionnaire is presented to the participant, the questions are alternated between one positive and one negative question so that the participant needs to read every question, thus removing overall bias.

Furthermore, Brooke explains that the SUS outputs a single numerical value that allows an expert to compare the usability of a system with other systems [16]. For calculating the SUS score, the following method should be used [16]: First, add up the total score for odd-numbered questions and subtract 5 from the result to obtain x . Second, add up the total score for even-numbered questions and subtract that number from 25 to obtain y . Finally, add up the values of x and y and multiply the result by 2.5. [16]

$$SUS = 2.5 \times [(\sum(OddQuestionScores) - 5) + (25 - \sum(EvenQuestionScores))] \quad (3.1)$$

[16]

In a review of the method done in 2013, Brooke [15] explains that the fact that the scale is multiplied by 2.5 is just a *marketing* decision to make the scale more readable and more intuitive since it now has a maximum value of 100. That decision led to many people confusing the SUS with a percentage scale, which is a mistake [92].

However, the SUS is still a very useful method, and several academics have identified that it represents a true perception that the users may have of a system [60] [59] [55]. Furthermore, in 2009 Bangor et al. [11] studied the results of different evaluations using

the SUS and developed acceptability rates within the scale. Their adjective ratings can be seen in figure 3.2.

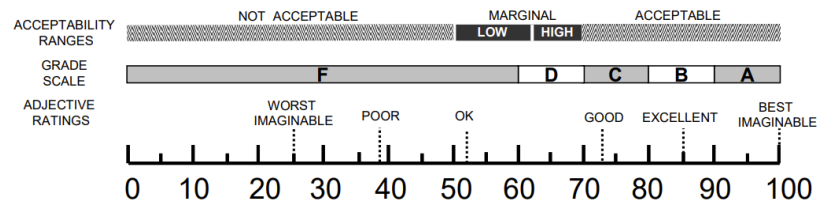


Figure 3.2: System Usability Scale and Adjective ratings. Developed by Bangor et al. in 2009 [11]

Despite being a quick and widely used method, the SUS has also been criticised by some researchers. For instance, Sauro and Lewis [103] criticised that the method required one positive and one negative question, one after the other. They explain that they found no evidence that providing mixed questions would allow users to provide more honest feedback. They found that intercalated questions may increase the amount of wrongly-labelled questions.

3.2.4 Questionnaires

Questionnaires are a method of gathering information in which a series of questions are asked to a participant. As cited by Kuter and Yilmaz [56], Brehob defined a questionnaire as “a form that people fill out, used to obtain demographic information and views and interests of those questioned” [14]. In the field of usability research, questionnaires are often used for obtaining more structured information, which eases the procedure of data analysis.

Questionnaires often output structured data, which can be easy to analyse and interpret. For that reason, several examples can be found on the literature of researchers using questionnaires alongside Think Aloud [28], [104], [120].

3.3 Related Work

As stated before, usability testing methods often output unstructured data, which is tedious and costly to process. Several proposals have been made before to try and overcome such issues. For instance, Soure et al. proposed CoUX [106], a tool that allows an expert to analyse a Think Aloud [57] session by checking a video of the

participant session. As stated by the authors, CoUX automatically detects sentiments, which the expert can use for annotating different parts of the interface. As can be seen from the GitHub repository for CoUX [52], it is not deployed for public usage but is open source and available for download.

Another example can be found in work done by Chanchí et al. [17], who proposed in 2018 a tool that automatically extracts data from an emotional analysis of the voice of a participant's Think Aloud [57] session. They use the *circumplex model of affect* proposed by Russell [101] to detect the emotions in the participant's voice and claim that it allows them to estimate a user's satisfaction when using a system. However, their work has no insights on being deployed, so it remains academic research.

Apart from the academia, many tools can be found online to assess the usability of a system. Examples of those platforms are Maze [66], UserTesting [116], TryMyUi [111], Hotjar [48] or Google Analytics [40], which can help design teams to evaluate their systems at different maturity stages. However, the market for usability research tools is very broad and there are different tools for different research objectives and budgets.

Additionally, as stated by Firmenich et al. [34], many companies avoid usability evaluation altogether due to its high costs, especially small and medium-sized companies. That is why it is vital to review the market of those tools from an economic point of view to understand how the companies could adopt a new tool.

After a look-up on different specialised forums and search engines, I built a list of fifteen tools that could be used for assessing the usability of a website. Since most of the tools are aimed at different evaluation objectives, I defined common criteria to filter the tools that are compatible to the objective of this project. As a result, I identified five tools that implement three fundamental concepts of remote usability testing with users: first, they all allow an expert to set up tasks on a website or prototype. Second, they can recruit participants and those participants can execute the tasks remotely. Finally, experts can analyse the results after the participants finish.

The tools that met the criteria for analysis were UserTesting [116], Maze [66], UxTweak [118], TryMyUi [111] and Lookback [62]. The descriptions of their main features can be found in the appendix A.2 and the benchmark analysis of their features can be found in the following chapter, in section 4.5.

Chapter 4

Requirement gathering and design

In this chapter, I present the process of gathering the requirements for the tool by interviewing five HCI experts from The University of Edinburgh. This project phase was carried out using a user study approved by The University of Edinburgh School of Informatics under the code 2022/61691.

4.1 Aims and Objectives

This study aims to obtain the final pieces of information for answering research questions RQ1 and RQ2, as well as answering research question RQ3.

4.2 Protocol

Following the University of Edinburgh's ethics procedure, I contacted five experts from the School of Informatics and invited them to participate in one-to-one meetings via Microsoft Teams [108]. During the 30-minute meetings, I used a semi-structured interview divided into two sections following the interview guide that can be found in the appendix B.4.

I designed the interview guide with two principal objectives. First, I wanted to obtain knowledge about how Think Aloud [57] and Question-Asking Protocol [54] are used in real studies. Second, I aimed to gather knowledge on how those methods can be integrated into a tool. Therefore, I first included questions regarding the issues the experts had using those methods and in which cases they recommended their usage. I also included questions about the best practices while using both methods and how they prepare for them.

Second, I included questions about specific features I thought could be included. Those functionalities were obtained after an initial hand-drawn sketch that I did, based on my experiences using the five tools mentioned in chapter 3.3. It was a very early draft that can be found in the appendix B.1, but it allowed me to identify which features I had to ask the experts about. Furthermore, I saw that only one of the tools I reviewed explicitly allowed Think Aloud, and none of them mentioned Question-Asking Protocol despite mentioning other methods such as the System Usability Scale [16]. For that reason, I also included questions about the advantages and disadvantages of both methods, as I wanted to understand if there is an underlying reason why they are not widely marketed in the existing online tools.

Semi-structured interviews were the appropriate method for this evaluation because all the participants in the study had different experiences and opinions, so I could ask further questions and comments on the parts I found more interesting. The evaluation produced more than 2 hours of video, which was stored in the University's Microsoft SharePoint [70] server, guaranteeing confidentiality and data protection.

More information about the participant recruitment, materials and further procedure can be found in the appendix B.

4.3 Data analysis

Data analysis for the requirement gathering phase focused on qualitative analysis of the transcripts obtained from the interviews. I processed the transcripts of all the interviews using NVivo [87] and conducted a Thematic Analysis [19] on the answers given by the experts. I chose Thematic Analysis because it allowed me to categorise the different answers the experts gave.

I used NVivo [87] to split the data into different topics and defined different theme hierarchies. I began by splitting the themes into the main groups of questions I asked. I defined six themes: Question-Asking Protocol, Think Aloud [57], evaluation preparation, failed tasks, after-evaluation procedure and tool features. Inside those themes, I also included different sub-themes, which evolved as I analysed more interviews. I included sub-themes such as the advantages and disadvantages for Think Aloud [57] or the different categories of features for Wanda.

Figures on the theme hierarchy and definitions can be found in the appendix B.6.

4.4 Results

Overall the five interviewed experts provided great feedback on their experiences with Think Aloud [57] and Question-Asking Protocol [54]. In general terms, they expressed many advantages for both methods, such as the fact that it is fast to execute, that it provides excellent insights on the usability of the evaluated system and that it is *effective* for understanding usability issues. They were, however, very clear about the drawbacks of both methods. For example, they mostly agreed that those methods could be time-consuming and stressful for participants, who may feel watched and judged.

The experts also provided excellent insights on what Wanda should look like and which features it must include. After evaluating and contrasting the experts' feedback with the tools reviewed in chapter 3.3, I built a list of 23 user stories containing the features developed in the subsequent phases of this project.

First, experts mostly agreed that scripts are a crucial part of these kinds of evaluations with users, so they should have a way of setting up and reusing scripts during the different evaluations. They said it is essential to define the research question and the tasks for the evaluation, and they also suggested that the system include questionnaires after the Think Aloud sessions [57]. Two of the five experts also mentioned that the System Usability Scale [16] should also be incorporated, as it allows for quick comparisons between different iterations of a system.

Regarding the execution of studies with users, experts agreed that the most challenging part of Think Aloud is data analysis. They mentioned that it is essential to succeed with Think Aloud [57] to split the participant audio by each task and to have automatic transcriptions that can be exported to analyse data faster.

Three experts also mentioned that it would be good to have an interface that automatically calculates insights on task completion and/or task success rates amongst all participants, to identify which tasks are more complicated. Such features are available in some of the tools mentioned before. However, when queried about which tools they had used before, none of the experts mentioned the five tools I had evaluated.

Finally, only one of the five experts interviewed had previous experience with the Question-Asking Protocol [54] method. They recommended adding a feature for question classification and also looking at the *question-suggestion* protocol [43] in which the expert can also ask questions.

More detail on the experts' insights and responses can be found in the appendix B.7.

4.5 Discussion

4.5.1 Real life challenges of Think Aloud and Question-Asking Protocol

From the evidence obtained by the interviews conducted with the usability experts and contrasting it with the literature, I was able to conclude that the main challenges faced by usability researchers while conducting evaluations using Think Aloud [57] can be summarised into three issues, therefore answering the research question RQ1.

First, the Think Aloud [57] is a method based on a participant speaking aloud all the thoughts, so it is prone to producing vast amounts of unstructured data, which can take very long to analyse. This issue was expressed by three of the interviewed experts, and that was also studied by Soure et al. [106].

Second, two experts expressed that Think Aloud [57] is a method that may seem unnatural or uncomfortable for the participants, which is an issue that was also explained by Alhadreti et al. [3]. The fact that participants need to verbalise their thoughts already puts them in a situation they are not used to, which may be difficult for some participants and harm the amount of feedback they can give.

Finally, it can be inferred from the answers given by the experts that the penetration of usability research tools is not broad, at least in academic research. All the experts were queried about tools they used to execute Think Aloud [57] evaluations, and none of the evaluated tools in section 3.3 was mentioned by them. The fact that the experts did not mention any of the tools may imply that there are no tools that meet their needs, but five interviewees is a very small sample size to draw any conclusions.

As an additional note, it is important to express that the Question-Asking Protocol is not widely used. It can be concluded from the literature review, market research and interviews that the method is not very popular. None of the tools reviewed mentions the method, and only one of the five experts interviewed mentioned having used it. Therefore, this project did not provide sufficient data to answer the real-life challenges of the Question-Asking Protocol method.

4.5.2 Advantages and disadvantages of existing tools

The fact that experts did not mention the tools reviewed in section 3.3 does not mean that the tools do not fit their needs. In fact, some of the issues expressed by the experts and the features requested by them are already solved by some of those tools. That is why I

conducted a comparative analysis of the reviewed tools to identify their advantages and disadvantages and thus answer the research question RQ2.

A table with common evaluation criteria was defined to assess the tools objectively. The table considers five categories in three levels, depending on the number of features they provide or the price they charge. Those five criteria were chosen in relation to the most frequent features required by the experts. Furthermore, I defined three levels measuring how advanced each feature is in each tool. The table with the criteria is shown in figure 4.1.

		Level		
		1	2	3
Criteria	Website testing	does not allow testing on websites, only on design prototypes	allows websites but a script is required	allows a website
	Task definition	does not allow tasks or questionnaires	allows either tasks or questionnaires but are paid	allows tasks and questionnaires for free
	Participant recruitment	does not allow participant recruitment	allows participant recruitment but costs	allows participant recruitment for free
	Think Aloud	does not allow participant voice, screen or video recording	does not allow Think Aloud but records participants audio, screen or	Allows Think Aloud
	Data analysis	does not provide data analytics	provides data analytics but does not allow data export	provides data analytics and data export
	Cost	does not have a free tier	has a free tier	is either free or open-source

Figure 4.1: Criteria for evaluating existing usability research tools.

Figure 4.2 depicts the various features evaluated in the x axis. The figure shows the levels to which each tool can reach within each feature on the y axis. Each of the five tools under consideration is represented by a different colour.

All of the evaluated tools support participant recruitment and allow the expert to define tasks that participants can then execute. An expert can see some analytics within those tools, and some of the tools include dashboards with data on task completion. Furthermore, most tools enable the expert to replay either the audio or video recording of the participant session, and some even provide automatic audio transcription.

In one of the evaluated tools, it was impossible to include a deployed website, as they only allowed design prototypes. From the tools that allowed production websites, either a script, an additional program or a browser extension was required for conducting the evaluation, which may cause barriers for some researchers.

Furthermore, all of the evaluated tools included participant recruitment systems. Most of them claim to have a participant base, which experts can choose to pay for in order to reach a larger testing audience. However, it is important to note that 4/5 of the evaluated tools provided some free tier credits, which an expert can use to determine if the tool is appropriate for their needs. Furthermore, only one of the tools

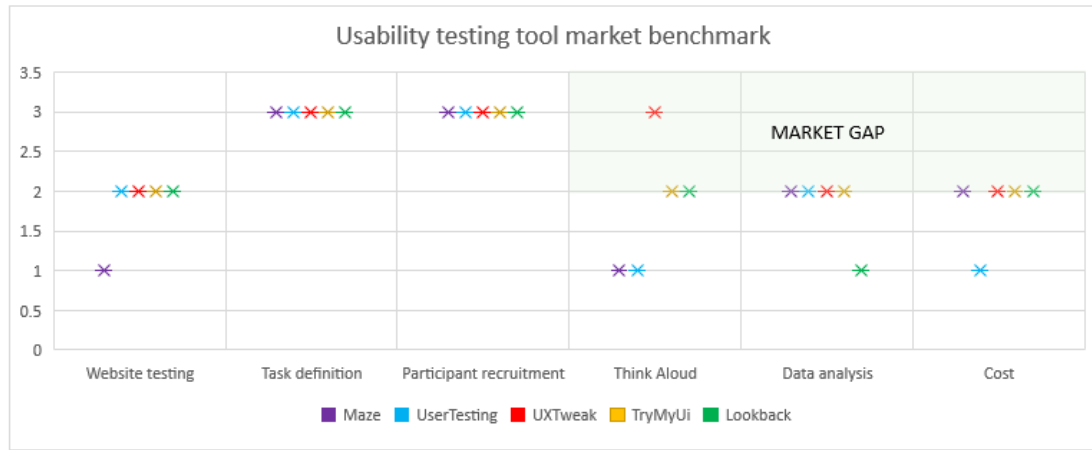


Figure 4.2: Benchmark of some usability evaluation tools available on the market.

evaluated explicitly recommended the Think Aloud method, but all of them provided the functionality of having questionnaires at the end of the session.

Finally, after searching for the tools on GitHub [38], one can see that none of them are open-source, indicating a clear market gap. As a result of the literature review, related work, and requirement gathering phases of this project, we can conclude that no widely used open-source tool supports task definition, participant recruitment, and data analysis and export features with the Think Aloud and Question-Asking Protocol methods.

4.5.3 Design of Wanda

To answer the research question RQ3, I summarised all of the features found on existing tools and compared them to the features requested by the experts in the interviews. I concluded that experts need an easy-to-use tool to design, conduct, and analyse usability evaluations. Some tools on the market already have those features, but they all come at a price that not every company or researcher can afford. As a result, one of Wanda's main requirements is to be open-source, so that any company or researcher can clone and deploy it, saving money and perceiving the benefits of usability research.

In conclusion, I built a list of 27 user stories, 23 of which were prioritised and can be found in the appendix B.8. Furthermore, the design of Wanda needs to include three categories of features: First, functionalities for quickly setting up an evaluation on a website. Second, the ability to invite participants to a study and the ability for them to conduct it. Finally, features for data export and analysis.

Chapter 5

First iteration of development

In this chapter, I present two topics: first, the major technical decisions made on Wanda based on the requirement gathering. Second, I present the main features developed during this stage of the project. The process followed in this iteration allowed me to start building an answer to the research question RQ4.

5.1 Development stack

A full-stack web application can be built with a variety of technologies. For example, an application could be written in Python [96], Java [51], .NET [71], or Node.JS[83], amongst others. I chose Node.JS because it is an engine in which I have prior experience, so there is almost no learning curve, which is critical for meeting project deadlines.

Wanda was built on Next.js [77], a React.js [98] and Node.js [83] framework. Amongst the various Node.js frameworks available, I selected Next.js because it is a production-ready environment with simple deployments to *Platforms as a Service (PaaS)* like Vercel [122], Netlify [75] or Heroku, from which I chose Vercel because it is the one I have experience with. Furthermore, Next.js is appropriate for this project because it is a single repository where both the front-end and back-end can be developed, reducing developer's need to maintain specific infrastructure.

The system uses a relational database hosted on AWS's [9] RDS service, where a PostgreSQL [93] cluster was created. A relational database was preferred over a NoSQL database because this application requires a lot of data manipulation, which is easier to do with a relational database supporting the SQL language [67].

Wanda comprises one materialized view and 16 tables that enable the whole functionality. The description of all the tables can be found in the appendices C.2.1 and

C.2.2. For creating those tables and managing database communications, I implemented Prisma, an Object Relational Mapper (ORM) library that was chosen amongst other ORMs like TypeORM due to its simplicity and robustness.

5.2 Back-end

Next.js [77] allows back-end code execution leveraging what they call *API routes*. An API route is a *stateless* function deployed to a Lambda [10] serverless environment that can execute back-end procedures. For defining an API route, a new file is created under the */pages/api* folder. Whenever a request is made to those API routes, a new Lambda function is executed and responds to the request. Since those codes are executed in Lambda functions and are not exposed to the user's browser, it is safe to include back-end code and database communications.

For exposing the database resources to the back-end, a GraphQL [42] server was implemented, using the package *apollo-server-micro* [6]. The route */api/graphql* was created as the entry point for receiving all the GraphQL queries from the front-end. Every request to the route requires an active *session*, so the user must be authenticated before requesting resources from the the back-end.

GraphQL [42] was chosen for this project because, as opposed to REST APIs, it does not have issues with over-fetching [30]. Since this application requires frequent reads and writes to a database, it is crucial to maintain HTTP communications as small as possible to maintain the system fast and responsive. Since GraphQL allows the developer to select which fields are required by a query and can also fetch data from multiple tables in one query, it makes the interface fast and with short loading times.

The package Apollo Client [5] was used to communicate to the API on the client side because it is the most widely-adopted GraphQL [42] package for NodeJS [83]. Apollo client allows straightforward communication between the front-end and the back-end by its implementation of the hook *useQuery*. Apollo, like any other GraphQL [42] library, requires the definition of types, queries and mutations, which can be a tedious and error-prone task. Therefore, I implemented an open-source package called Prisma-Cosmo [25], which I previously developed for other projects. Prisma-Cosmo takes the Prisma Schema and automatically generates types and resolvers that the Apollo Server automatically picks up to enable the create, read, update and delete (CRUD) functionalities. There are other tools for generating automatic CRUD from a Prisma schema, like TypeGraphQL [112], but Prisma-Cosmo is lighter and faster to implement.

Furthermore, the NodeJS [83] ecosystem contains many authentication packages, such as PassportJS [90], NextAuth [76], or Auth0 [8], amongst others. I chose NextAuth over the other options because it integrates more easily with a NextJS [77] stack. It also supports multiple authentication providers, all of which are based on the OAuth2 [44] standard. More information on authentication is available in the appendix C.2.

5.3 Front-end

In this section, I provide explanations on the most important features developed in this iteration. When relevant, a feature is associated with its user story in parentheses. All the photos can be found in the appendix C.3.

5.3.1 Expert Interface

When experts log into the system, they can find three menus: One for setting up the scripts, one for setting up the evaluation studies and one for setting up the study sessions with their participants.

For creating a *script*, experts can open the script creation menu, which allows them to write all the script's content. Experts can set up different text formats for the script, such as quotes or links (1d). The expert must also record themselves reading the script because the system allows a participant to read or listen to the description (2c). The audio recording is mandatory because during the requirement gathering, it was suggested by one expert and it is a feature that can improve the accessibility of the tool.

Once the script is created, the expert can proceed to create an *evaluation study*. For doing so, the expert must define the name of the study (1b), the URL of the website to evaluate (1b), the target number of participants (1b) and the research question (1c). The expert can also select which script to use for the study (1d). The system requires the expert to define the tasks and the questionnaire for the evaluation (1a, 1e). For defining the tasks, the expert must input the URL and description of each task, and they must also record themselves reading the task, so that participants can later listen to the expert's voice instead of reading the task description.

For defining the questionnaire, the expert must first decide whether or not they want to use the System Usability Scale [16] (4d). Then, if the expert decides to use it, the system loads the ten standard questions of the SUS, but the expert can modify them if they want. The expert can also set up open-ended questions for the end of the session.

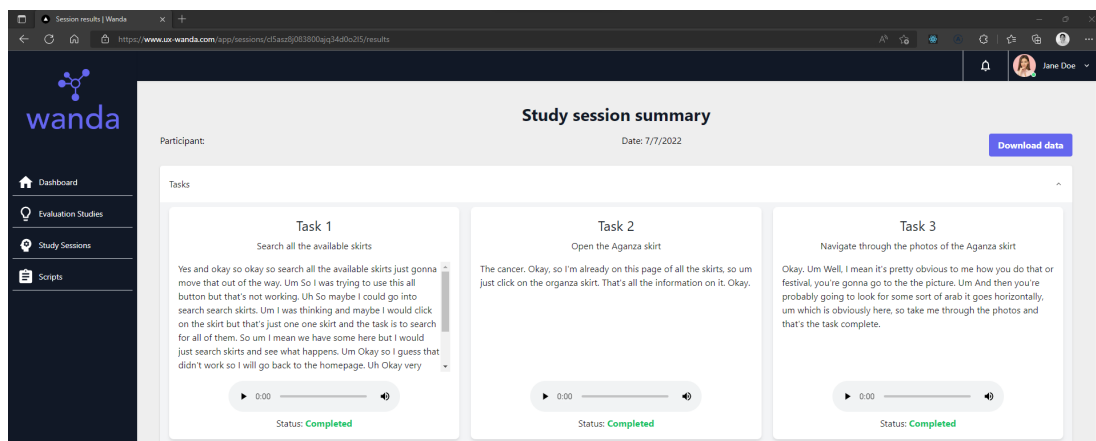


Figure 5.1: Experts' interface. First iteration of Wanda.

Once the evaluation session is created, the expert can go to the *Study sessions* menu to create a session with a participant. When the expert arrives at the study sessions interface, they see a list of all the studies conducted and can create a new one. To create a new session, they must input the participant's e-mail and select the evaluation skirt study in which the participant will participate (2a). By this iteration of Wanda, the participant is not receiving an automated e-mail when the session is created, so they need to log in manually. However, the system checks if the desired participant already has an account. If not, the system creates an account for the participant, which allows them to sign-in by sending them an e-mail that contains a magic link.

During the participant's session, the expert also has an interface where they can mark a task as completed or failed (2g). Once the evaluation finishes, experts can view its results (4c). The expert can see all the tasks and hear the participant's speech. The system automatically transcribes each task's audio from the participant (4a) and saves the text to the database, so the expert can also read what the participant said in each task (4b), allowing for faster data analysis. A sample photo of the session result interface can be seen in figure 5.1. Experts can also see all the individual answers for the questionnaire and a chart with the result for the System Usability Scale [16] (4c). Furthermore, the expert can to export the data for the session as JSON [49] (4e).

Finally, an expert can also see the results for an evaluation, aggregating all the sessions. When experts go to the page to see the evaluation results, they see a chart that presents each task's success rate and average duration (4c). The system also shows the average SUS [16] (4d) score aggregated by all the evaluation participants. The expert can also export the evaluation-wide data in Excel [68] format (4e). It is worth noting that Wanda contains more charts, that were not explicitly suggested during the requirement

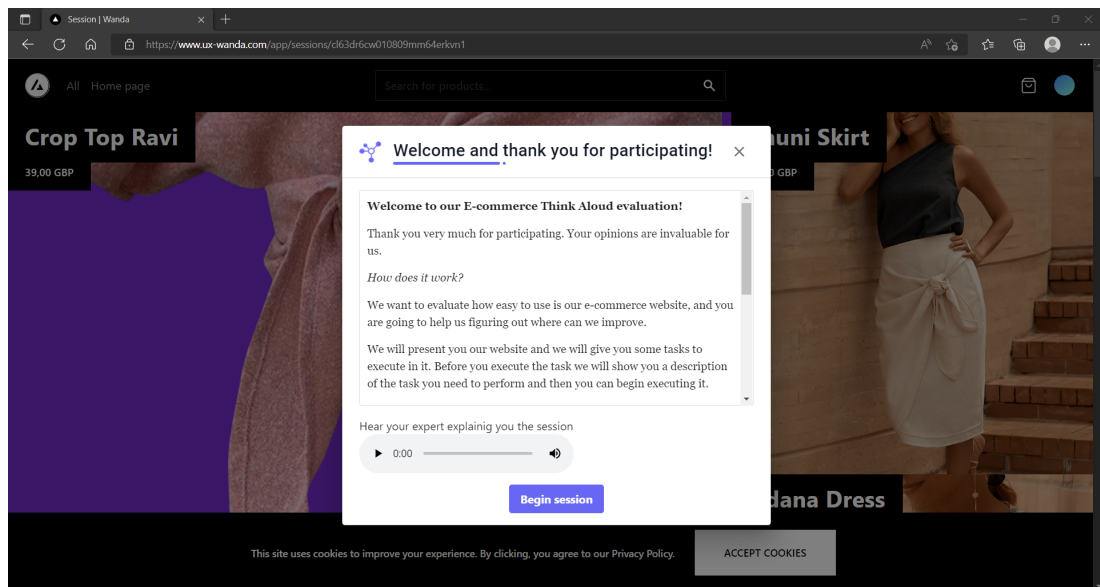


Figure 5.2: Participants' interface. First iteration of Wanda.

gathering. However, I decided to include them because all the tools reviewed included some sort of dynamic visualisation of data.

5.3.2 Participant Interface

When participants log in, they see a table listing all of the study sessions to which they have been invited. Each row in the table represents a study session, which the participant can start by clicking the *launch* icon (2a). When a participant begins a session (2b), the system displays a window with the expert's previously configured script, which the participant can read or listen to (2c). When the participant has finished reading the script and is ready to start, they can click *begin session*, and the system will present the first task (2d), which they can also read or hear (2e). The system records their audio as they complete the task. When participants believe the task is complete, they can use the *task controls* to mark the task as completed or failed (2f). When a task is completed, and the expert confirms its status, the system automatically presents the next task, and the process is repeated until all tasks are completed. A sample photo the participants' interface can be seen in figure 5.2.

When participants finish the session, they can complete the questionnaire that the expert previously set up. Participants can navigate through the questions and answer them by selecting the option or typing the answer (2h). Once they finish answering all the questions, they can see a summary of all the answers and confirm them (2i).

Chapter 6

Formative evaluation

In this chapter I present the process of evaluating the impact of the tool formatively by interviewing HCI experts and students from The University of Edinburgh. This study was approved by The University of Edinburgh School of Informatics under the code 2022/61691. The study is divided into two sections: one for evaluating the interface for the experts and another one for evaluating the interface for the participants.

6.1 Aims and Objectives

This study aims to gather an initial feedback on both the participant and expert interfaces described previously on chapter 5. The feedback obtained in this study will also be used for building an answer to the research question RQ5a.

6.2 Protocol

For this evaluation we followed a similar procedure with the one seen on chapter 4. I contacted three experts in usability and ten students from The University of Edinburgh. I conducted one-to-one sessions with each of them through Microsoft Teams [108]. On the one hand, the sessions with the experts used Think Aloud [57], and they had to execute some tasks directly in Wanda for setting up an evaluation study on a mock e-commerce platform that I had previously set up and deployed. During each task, the experts gave me feedback on the functionalities of Wanda, and after they completed each task, I asked them questions on a semi-structured interview.

On the other hand, the students assumed the roles of participants on a study session on the aforementioned e-commerce platform. They were asked to log in to Wanda

and go to the session that was created for them. They had to execute four tasks on the e-commerce platform and answer a questionnaire about it. After they finished executing the study session, I asked them questions about their experiences as part of a semi-structured interview.

In both cases, semi-structured interviews were chosen because they allowed me to modify the questions depending on what I saw from their interactions. Furthermore, after the semi-structured interviews, I asked all the participants to complete a questionnaire for obtaining the System Usability Score [16].

More information about the recruitment, data collection methods, materials and procedure can be found in the appendix D.

6.3 Data analysis

Data analysis for the formative evaluation phase focused on qualitative analysis of the transcripts obtained from the one-to-one sessions. Similar to the requirement gathering presented in section 4, I also used Thematic Analysis [19] in NVivo [87] for analysing the results. However, since this evaluation considers two different use cases of Wanda, I followed a different strategy for defining the themes.

Even if the themes expressed by the experts and the participants were similar, I decided to keep them separate. The experts' and participants' themes include what they liked and disliked and the topic of new feature suggestions. However, during the evaluations, the experts faced more technical issues on the platform, so I decided to include a topic for them on the bugs they experienced. The participants had more broad comments on Wanda, so I included specific themes for the task controls, the questionnaire, the login system, the script and the question I directly asked them about how they would feel about using Wanda without an expert being present.

More detail on the experts' and participants' insights and responses can be found in the appendix D.5.

6.4 Results

The sessions conducted in this formative evaluation were very insightful. Overall, both experts and participants were satisfied with Wanda's interface and functionalities. On the participants' side, they used adjectives like *professional*, *intuitive* or *easy* to refer to their experience using Wanda. Naturally, there were issues with the system, and they

recommended some features to be considered for implementation. For instance, the most commented feature (6/10) was to include a tutorial at the beginning of the session to teach new users how the study session works. Furthermore, 9/10 of participants said they felt they could do the whole study on their own without the help of an expert.

On the experts' side, comments were also positive. They all agree that the most liked features were the data export system and the fact that Wanda has automatic transcriptions and that they are split by each of the tasks, which is aligned to what I presented in chapter 4. By the time of this evaluation, Wanda did not include responsive layouts, and all the experts used the system with minimised windows, which caused all of them to complain about the layout since it had components on top of other components, making the interface confusing. Additionally, 2/3 experts recommended changing the navigation through the menus to make them more straightforward. Furthermore, also 2/3 experts suggested changing the visualisations for the System Usability Scale [16] because the one presented is not the industry standard.

The participants gave Wanda a score of 87.2 and the experts a score of 81.7 on the System Usability Scale [16], which means that both user groups believe that Wanda has an *acceptable* usability. More detail on the experts' and participants' insights and responses can be found in the appendix D.6.

As a result for the formative evaluation, I built an updated list of ten requirements that served as input for the second iteration of the development. The list can be found in the appendix D.7.

6.5 Discussion

At this stage of the project, it was still early to draw conclusions on the research question RQ5. However, this formative evaluation provided very good feedback on Wanda's usability. The fact that participants felt that they would be available to do the evaluation on their own encouraged me to include that feature in the second iteration, as will be seen in chapter 7.

The feedback of the three experts is also very positive. Even though they experienced issues with the responsiveness of the platform and one expert did not find clear how to navigate through the menus, they were all successful in using Wanda for creating the evaluation study, which already shows that the tool may have potential in the future.

Chapter 7

Second iteration of development

In this chapter, I present the most significant changes made to Wanda after the feedback obtained in the formative evaluation. Most of the changes were front-end-related, and the application did not require any modifications to the backend or the infrastructure. I also provide the summary and answer to the research question RQ4. Throughout this chapter I will present different changes made to the first iteration, adding in parentheses the requirement to which the feature is associated. Additionally, I present all the pictures for this section in the appendix E.

7.1 Participant Interface

As explained in the previous chapter (6.4), the biggest complaint of the participants was that the system did not include an onboarding tutorial. For that reason, a tutorial was implemented to help participants to understand how the task flow works (4). The tutorial guides the participant through the first screens of the application and teaches them the different functionalities.

This iteration of Wanda also included a significant change in how participants connect to a session. As will be detailed in the following subsection, when an expert creates a new study session, the participant receives an e-mail with a link that signs them into the system and takes them directly to the evaluation they need to perform, so the participant no longer needs to go to the system and manually log in. This is a requirement aligned with the user story 3a, which was not implemented in the requirement gathering and therefore was included as an additional requirement (10).

Furthermore, considering that participants mostly agreed that they would be able to conduct the evaluation independently, the system now asks the expert if they want

to be present in the session with the participant. If they decide it is a participant-only session, then the system will not require expert approval to begin the session (2), which means that the participants can go on their own and conduct the evaluation. This feature is also linked to the fact that previously, the only method available was Think Aloud [57]. In this iteration, I added the option for the expert to decide if they want to execute a Question-Asking Protocol [54] session instead (1). It is important to note that the participants do not see a big difference between both methods, because for them the only change is in the instructions.

7.2 Expert Interface

The most significant complaint from the experts was that the system was not responsive. That is why I implemented different layouts for mobile screens, tablets and laptops, making sure that the system behaves as expected regardless the size of the window (6).

Second, some of the experts were confused about the flow within Wanda. For that reason, I implemented the suggestion made by one of the experts to change the sidebar menu options 7. The new pages are configured as follows: First, experts need to *design* a study. Second, once the study is designed, they can go and *conduct* a session with a participant. Finally, once the studies are done, the experts *analyse* the results.

Third, as explained above, experts can now define if they want to use Think Aloud [57] or Question-Asking Protocol [54]. They can also decide if they want to participate in the session with the participant or not. If they decide not to be present with the participant, then the participant needs to begin the session on their own (1 and 2). If experts decide to use Question-Asking Protocol, then after the evaluation they see an interface where they drag-and-drop each of the participants' questions and categorise them into the different targets of intentions proposed by [54]. A sample of the interface for the Question-Asking Protocol classification can be seen in 7.1.

Furthermore, experts also recommended that the SUS [16] be presented following Bangor's [11] proposal for displaying the SUS. After an initial search on NPM [86], I could conclude that there were no publicly available React components for showing the System Usability Scale. To fulfil this requirement (3), I developed an interactive React component based on [126] displaying the SUS and Bangor's [11] adjective ratings. As a side result, I also published this code on both GitHub [38], and NPM [86] as an independent package, which can now be used by any React developer wanting to display an interactive system usability scale on their website [26].

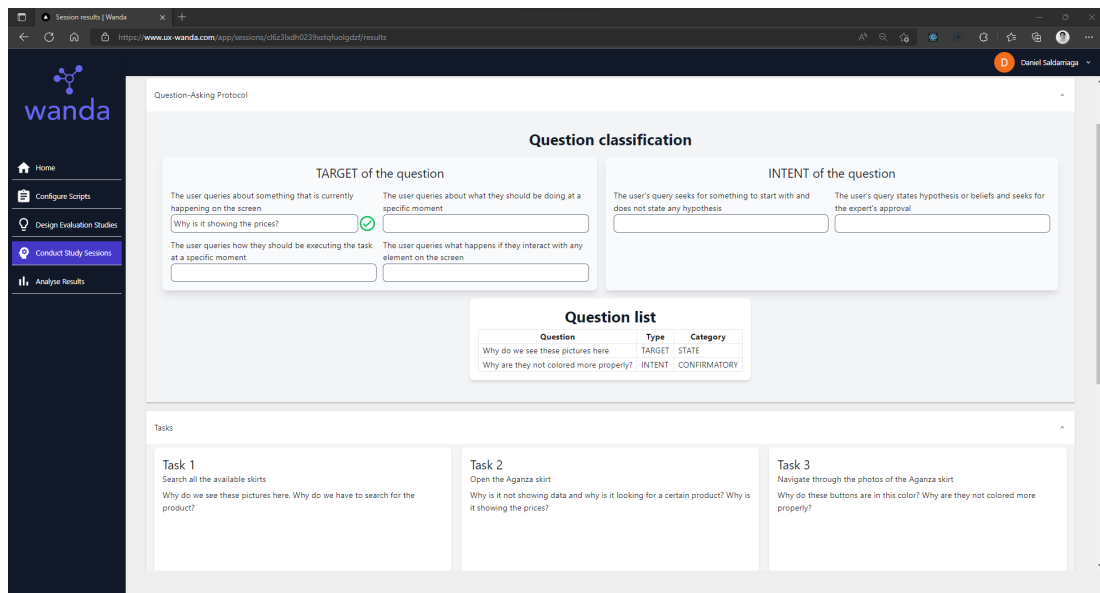


Figure 7.1: Question-Asking Protocol classification interface

7.3 Discussion

The development of a tool such as Wanda must be carefully planned and executed because it has many components that interact with each other and must be compiled into a system that has two different interfaces for two different types of users. Throughout the chapters 6, 5, and 7 I described the details of how I developed Wanda, allowing me to provide an insight on the research question RQ4.

The development of a complex tool is not something that can easily be done, but throughout this project, I was able to finish a working prototype that can already be used for experts to design, conduct and minimally analyse evaluation studies. So far, Wanda has been used by more than 13 people, providing valuable feedback on its features.

Additionally, I was able to assert the importance of doing a formative evaluation because the requirements given by an expert may not be enough to develop a system that can actually be used. For instance, doing one-to-one meetings with potential participants was the biggest source of fixes and new features that made it possible to develop the second iteration of Wanda.

Finally, analysing the results from the requirement gathering (4), the first iteration of development (5), the formative evaluation (6) and the results presented in the section above, I can conclude that it is technically possible to develop an open-source tool that allows an expert to create, conduct and analyse usability evaluation studies using the Think Aloud [57] and Question-Asking Protocol [54] methods.

Chapter 8

Summative evaluation

In this chapter I present the process of evaluating the impact of Wanda by interviewing HCI experts and students from The University of Edinburgh. This study was approved by The University of Edinburgh School of Informatics under the code 349375.

8.1 Aims and Objectives

This study aims to understand what do users believe about Wanda's final iteration of development, providing a summative view of the whole development process and answering the research question RQ5.

8.2 Protocol

The summative evaluation followed a similar procedure to the requirement gathering and the formative evaluation. I contacted five experts in usability and fourteen students from The University of Edinburgh, who were invited to one-to-one sessions and three group meetings, respectively.

Like the formative evaluation, all the sessions used Microsoft Teams [108]. The sessions with the experts used Think Aloud [57], and they had to execute some tasks directly in Wanda for designing, conducting and analysing the results of an evaluation study done on the same e-commerce platform explained before in chapter 6. During each task, the experts gave me feedback on the functionalities of Wanda, and after they completed each task, I asked them questions in a semi-structured interview.

Participants were invited to group meetings rather than one-on-one meetings because, due to time constraints, analysing three sessions rather than fourteen is faster. They

were instructed to log in to Wanda and go to the session created for them. I asked them to follow the tutorial and complete 4 tasks in the same mock e-commerce platform used in previous evaluations. However, I made sure to stop them before they completed each task and asked them questions about Wanda's features. When there were opposing viewpoints, I encouraged them to discuss them with the other participants. While they were executing their tasks, I asked them to close their microphones in Microsoft Teams [108] so that Wanda could take over the control of it and record their audio.

In addition, at the end of the sessions, I asked all participants to complete a questionnaire in order to obtain the System Usability Score [16]. I also asked participants to complete a questionnaire about Wanda's impact, useful for answering research question RQ5.

8.3 Data analysis

Like the previous evaluations, I used the strategy of leveraging Thematic Analysis [19] in NVivo [87]. I also separated the themes into those of the participants and those of the experts, even though they sometimes overlapped. However, unlike the formative evaluation, I decided to split the themes in each of Wanda's functionalities because I wanted to understand the impact of each functionality separately. The description and hierarchy of the themes can be seen in the appendix F.5.

8.4 Results

Overall, the feedback obtained in the summative evaluation was positive. All the 14 participants successfully completed all of the tasks assigned to them. The initial tutorial, which they all found simple to follow, received the most positive feedback. All of the participants also agreed that the log-in system was simple to use. However, despite being able to complete all of the tasks successfully, two of them were unable to provide verbal feedback, so they were unable to execute the Think Aloud [57] method. When asked why, they stated that they did not realise Wanda was recording their voice because there was no clear indication of the recording status. Their complaint was supported by 7/14 participants, which was the most serious issue they found in Wanda.

Experts also provided positive feedback and were able to successfully complete all of the tasks assigned to them. The experts' favourite features were the automatic transcripts divided by task and the ability to download data for the evaluation study.

They all found the visualisation for the System Usability Scale [16] interesting, and three of them commented on the potential of it being open-source. However, experts have also expressed concern about the audio recording feature. Three of the five experts, like the participants, stated that they did not understand how to use the voice recording.

As previously stated, both participants and experts were asked to complete the System Usability Scale [16] for Wanda. The participants gave a score of 81, while the experts scored 94. Both results are still *acceptable*, but it is worth noting that the score on the participants' interface has dropped from 87.2 to 81, probably due to having different use cases between both evaluations, as explained in section 9.1.1. On the experts' interface, the SUS increased from 81.7 to 94, evidencing a good reception of the changes made during the second iteration.

Finally, all both participants and experts were asked about Wanda's potential impact. The participants believe that Wanda's most significant potential is that it makes them feel confident while providing feedback on the evaluated system, as shown in figure F.5. Furthermore, experts believe Wanda's most significant potential is that it helps them analyse data more quickly, as shown in figure F.7. More insights on the answers for the participants and experts can be seen in the appendix F.6.

8.5 Discussion

In conclusion, this evaluation helped me understand that Wanda received positive feedback from both experts and participants. The fact that experts were able to successfully design, conduct, and analyse the results of a usability evaluation indicates that Wanda is a minimum viable project that could be useful if further developed and evaluated.

Regarding answers to the research question RQ5, I can conclude that Wanda has the potential for making the process of designing, conducting and analysing usability studies. However, as will be further exposed in section 9.1.1, a total of 28 people have participated in the studies of this project, which is not a sufficient amount of participants to draw definitive conclusions on Wanda's impact.

After two development iterations and formative and summative evaluations, I can conclude that Wanda is a viable tool for conducting usability studies and has the potential to gather honest feedback from participants. Despite the small number of interviewees, their overall comments were positive, and they felt confident while using Wanda, indicating that they can use the platform without major issues. This implies that, if developed further, Wanda could indeed be used for large-scale evaluation studies.

Chapter 9

Discussion, future work and conclusions

This chapter aims to summarise all the findings from this project, pointing out its limitations and possible future work.

9.1 Discussion

Overall, I think the findings and results of this project are quite interesting. I believe Wanda has the potential to assist experts in conducting distributed usability studies. The methodology used throughout the project taught me the value of specifying clear evaluation phases and comprehending the differences between evaluations at various stages of the project. It was a great exercise to compare the literature and existing tools to the features requested by the experts because I discovered that there are no widely-used open-source (or free) tools designed specifically for Think Aloud [57] and Question-Asking Protocol [54]. Additionally, I must attribute a big part of the success in the evaluations to the technological stack I chose for the project, because it allowed me to iterate fast and develop a robust array of functionalities.

Furthermore, I found that by making Wanda open-source, many companies that cannot afford to pay for the existing market tools could incorporate usability studies. In addition, as a side effect of the project, I published a second open-source package for displaying the System Usability Scale [26], which was not planned initially but was greatly appreciated by the experts.

I was also able to capitalise on the benefit of having both a formative and a summative evaluation as I developed features that were better suited to the needs of experts

and participants. It was an interesting exercise to see users request a feature during the formative evaluation and then see it implemented during the summative evaluation because they could see improvements in the platform's usability. Following the completion of both the formative and summative evaluations, I concluded that Wanda could be used easily by both participants and experts, implying that it has the potential to be widely adopted.

9.1.1 Limitations

This project had numerous limitations that the reader should be aware of. First, despite the fact that I conducted three evaluations for this project, the total number of participants was small. The three studies included nine usability experts and 19 participants. Twenty-eight participants is not a large enough sample size to accurately evaluate Wanda. Furthermore, because everyone involved in the evaluations for this study is from an academic background, the feedback obtained may be biased toward a research-oriented speech. A more extensive study with different enterprises and participants across different markets is required to assess Wanda's operational usefulness. As a result, more testing and evaluations are required to determine the tool's true impact.

Second, as previously stated, the SUS [16] for participants decreased from 87.4 to 81.7 between the formative and summative evaluations. However, neither evaluation used the same methodology. As seen in section 6.2, on the one hand, for the formative evaluation, participants were part of a one-on-one formative evaluation session, which may have caused them to provide less rough feedback because I was present in the session with them. In addition, I was the one who explained them how to use Wanda, and they did not need to learn it themselves. On the other hand, in the summative evaluation, participants used the system on their own, which may have caused them to struggle a bit more with the interface. The fact that they were different evaluations demonstrates that the SUS results are not comparable because they measure different use cases of Wanda.

Another limitation of this project is that on the summative evaluation, participants were explicitly queried about Wanda's impact. There is a problem with the questions because they are worded positively. As a result, participants may have been biased toward answering the questions positively, which means the results may be misleading. Finally, this project has a time constraint. Because this is a Master's dissertation project, it is limited to 8 weeks of development time and a single developer. Therefore, with

more time and resources, this project could have provided more features and a more robust evaluation procedure.

9.2 Future work

Both participants and experts complained about the recording feature implemented in Wanda, as seen in section 8.4. It was unclear to them when the system was actively recording their audio, so this feature should be addressed as a top priority. To handle this issue, I would suggest adding a draggable microphone icon that indicates whether or not Wanda is recording the audio similarly to how it is done in messaging apps.

Wanda, as previously stated, is a tool with promising potential. However, it is missing some features. The most important feature that was not implemented is the ability to record the screen and the participant's video. That is a feature that the experts requested during requirement gathering, and it was mentioned in both the formative and summative evaluations. It was not implemented due to time constraints, but the architecture was left prepared for receiving it. Wanda already makes use of JavaScript's Media Recorder API [72] for capturing audio, and it can also capture video and screen recordings, so it would be a matter of tweaking the code for activating those channels.

Finally, Wanda could leverage the same architecture and infrastructure for adding more evaluation methods. The modifications in the code for adding methods like cognitive walkthroughs [124] or formal inspections [53] would not be significant. A complete list of future requirements can be found in the appendix F.8.

9.3 Conclusions

During this project, I designed, developed, and tested an early version of Wanda, a tool that enables a usability expert to design, conduct, and analyse remote studies on websites using the Think Aloud [57] and Question-Asking Protocol [54] methods. The project began with a review of the literature and related work, with the goal of determining which tools on the market provide similar functionalities. After determining the market status, I conducted a requirement gathering phase in which I interviewed five usability research experts. With their feedback, I was able to create a list of requirements, which can be found at section B.8 and were used for the first iteration of Wanda.

Following the completion of the first iteration, I conducted a formative evaluation, which included one-on-one sessions with three usability experts and ten participants.

The formative evaluation enabled me to identify some Wanda features that needed to be improved and new features that were missing, so I developed a second iteration. Once it was completed, I conducted a summative evaluation, interviewing five usability experts and holding three focus groups with 14 participants.

This project had five research questions, which were addressed in different sections throughout this paper. First, using the related work review and requirement gathering, I could answer RQ1 in section 4.5.1. I concluded that the real-world problems with the Think Aloud [57] and Question-Asking [54] protocols are primarily based on three issues. The first is the lack of a tool designed specifically for such methods. The second issue is the high cost of existing solutions. Finally, the third issue is that those evaluations are currently carried out either in person or with tools designed for other purposes, making data analysis difficult.

I was able to answer RQ2 in section 4.5.2 by conducting related work research and contrasting it with requirement gathering. I determined that there is no widely used open-source tool specifically designed for Think Aloud [57] or Question-Asking Protocol [54]. I only found one tool that used Think Aloud [57], and it did not allow data export, so it did not solve the problem that experts face with tedious data analysis.

I was able to answer RQ3 after completing the requirement gathering phase. I created a list of 23 user stories B.8 that represented experts' needs and aimed to address the most pressing pain points they expressed. I came to the conclusion that experts need an easy-to-use tool for designing, conducting, and analysing remote usability studies.

Furthermore, after two development iterations, I was able to answer RQ4. I could conclude that creating an open-source tool for designing, executing, and analysing remote usability studies is viable on a technical and operational level. Even with the limitations stated in section 9.1.1, I can conclude that Wanda has promising potential and, if further developed and evaluated, can positively impact experts' pain points.

Similarly, following the summative evaluation, I could answer RQ5. First, based on the System Usability Scale [16] results, I concluded that the interviewed experts and participants believe Wanda has an *acceptable* usability level. Second, I could conclude that Wanda could potentially shorten the time it takes experts to analyse the findings of an evaluation study. Finally, based on the information gathered, I concluded that both experts and participants believe Wanda may be able to assist participants in providing more feedback on an evaluated site because it appears simple to use. However, as stated in section 9.1.1, only nine experts and 19 participants helped evaluating Wanda, so more iterations are needed to reach a definitive conclusion on Wanda's true impact.

Bibliography

- [1] ISO 9241-11: 2018. Ergonomics of human-system interaction—part 11: Usability: Definitions and concepts, 2018.
- [2] Adobe. Adobe XD: Fast and powerful UI/UX design and collaboration tool. <https://www.adobe.com/products/xd.html>.
- [3] Obead Alhadreti and Pam Mayhew. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.
- [4] Amazon. Amazon.com. <https://www.amazon.com/>.
- [5] Apollo Client. Comprehensive state management library for JavaScript that enables you to manage both local and remote data with GraphQL. <https://www.apollographql.com/docs/react/>.
- [6] Apollo Server Micro. Micro integration for the Apollo community GraphQL Server. <https://www.npmjs.com/package/apollo-server-micro>.
- [7] Apple. Safari - Apple. <https://www.apple.com/safari/>.
- [8] Auth0. Auth0: secure access for everyone. <https://auth0.com/>.
- [9] AWS. Cloud Computing Services. <https://aws.amazon.com/>.
- [10] AWS. Lambda: Run code without thinking about servers or clusters. https://aws.amazon.com/lambda/?nc1=h_ls.
- [11] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

- [12] Randolph G Bias. The pluralistic usability walkthrough: coordinated empathies. In *Usability inspection methods*, pages 63–76. 1994.
- [13] Randolph G. Bias and Clare-Marie Karat. 1 chapter - justifying cost-justifying usability. In Randolph G. Bias and Deborah J. Mayhew, editors, *Cost-Justifying Usability (Second Edition)*, Interactive Technologies, pages 1–16. Morgan Kaufmann, San Francisco, second edition edition, 2005.
- [14] K Brehob et al. Usability glossary. *Online*, <http://www.usabilityfirst.com>, [2 Sep. 2005], 2001.
- [15] John Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2):29–40, 2013.
- [16] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
- [17] Gabriel E Chanchí G, Luis F Muñoz S, and Wilmar Y Campo M. Proposal of a tool for the stimulation of satisfaction in usability test under the approach of thinking aloud. In *International Congress of Telematics and Computing*, pages 211–222. Springer, 2018.
- [18] Chi-Cheng Chang and Tristan Johnson. Integrating heuristics and think-aloud approach to evaluate the usability of game-based learning material. *Journal of Computers in Education*, 8(1):137–157, 2021.
- [19] Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. *Qualitative psychology: A practical guide to research methods*, 222(2015):248, 2015.
- [20] Aurora Constantin, Cristina Alexandru, Jessica Korte, Cara Wilson, Jerry Alan Fails, Gavin Sim, Janet C Read, and Eva Eriksson. Distributing participation in design: Addressing challenges of a global pandemic. *International Journal of Child-Computer Interaction*, 28:100255, 2021.
- [21] Michael E. Cope and Kevin C. Uliano. Cost-justifying usability engineering: A real world example. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 39(4):263–267, 1995.
- [22] Deborah Cotton and Karen Gresty. Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology*, 37(1):45–54, 2006.

- [23] CrazyEgg. Website optimisation, A/B testing and heatmaps. <https://www.crazyegg.com/>.
- [24] David Crow. 6 chapter - valuing usability for startups. In Randolph G. Bias and Deborah J. Mayhew, editors, *Cost-Justifying Usability (Second Edition)*, Interactive Technologies, pages 165–184. Morgan Kaufmann, San Francisco, second edition edition, 2005.
- [25] Daniel Saldarriaga. prisma-cosmo: automatic types and resolvers for basic CRUD from a Prisma Schema. <https://github.com/prevalentWare/prisma-cosmo>.
- [26] Daniel Saldarriaga. React component for displaying the System Usability Scale. <https://github.com/danyel117/react-system-usability-scale>.
- [27] George M Donahue, Susan Weinschenk, and Julie Nowicki. Usability is good business. *Compuware Corp., julio*, 1999.
- [28] Afke Donker and Panos Markopoulos. A comparison of think-aloud, questionnaires and interviews for testing usability with children. In *People and computers XVI-Memorable yet invisible*, pages 305–316. Springer, 2002.
- [29] ebay. Ebay. <https://www.ebay.co.uk/>.
- [30] Thomas Eizinger. Api design in distributed systems: a comparison between graphql and rest. *University of Applied Sciences Technikum Wien-Degree Program Software Engineering*, 2017.
- [31] Ellen Emanuela, Ari Widyanti, and Gradiyan Budi Pratama. Usability evaluation of a fintech lending mobile application for university student: A case study. In *IOP Conference Series: Materials Science and Engineering*, volume 1077, page 012058. IOP Publishing, 2021.
- [32] K Anders Ericsson and Herbert A Simon. Verbal reports as data. *Psychological review*, 87(3), 1980.
- [33] Figma. The collaborative interface design tool. <https://www.figma.com/>.
- [34] Sergio Firmenich, Alejandra Garrido, Julián Grigera, José Matías Rivero, and Gustavo Rossi. Usability improvement through a/b testing and refactoring. *Software Quality Journal*, 27(1):203–240, 2019.

- [35] FiveSecondTest. Five second tests. <https://fivesecondtest.com/>.
- [36] World Economic Forum and Boston Consulting Group. Internet for all: A framework for accelerating internet access and adoption, 2016.
- [37] Edwin Gamboa, Rahul Galda, Cindy Mayas, and Matthias Hirth. The crowd thinks aloud: Crowdsourcing usability testing with the thinking aloud method. In *International Conference on Human-Computer Interaction*, pages 24–39. Springer, 2021.
- [38] GitHub. GitHub: Where the world builds software. <https://github.com/>.
- [39] Google. Google Scholar. <https://scholar.google.com/>.
- [40] Google Analytics. Google Analytics: Know your audience. <https://analytics.google.com/analytics/web>.
- [41] Google Cloud Platform. Cloud Computing Services. <https://cloud.google.com>.
- [42] GraphQL. A query language for your API. <https://graphql.org/>.
- [43] Tovi Grossman, George Fitzmaurice, and Ramtin Attar. A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 649–658, 2009.
- [44] Dick Hardt. The oauth 2.0 authorization framework. Technical report, 2012.
- [45] H Rex Hartson, Terence S Andre, and Robert C Williges. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, 15(1):145–181, 2003.
- [46] Morten Hertzum, Kristin D Hansen, and Hans HK Andersen. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2):165–181, 2009.
- [47] Tasha Hollingsed and David G Novick. Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th annual ACM international conference on Design of communication*, pages 249–255, 2007.
- [48] Hotjar. Hotjar: website heatmaps & behaviour analytics tools. <https://www.hotjar.com/>.

- [49] Ecma International. Ecma-404—the json data interchange format, 2013.
- [50] Kashif Ishaq, Fadhilah Rosdi, Nor Azan Mat Zin, and Adnan Abid. Heuristics and think-aloud method for evaluating the usability of game-based language learning. *International Journal of Advanced Computer Science and Applications*, 12(11), 2021.
- [51] Java. Java programming language. <https://www.java.com/en/>.
- [52] Jian Zhao, Emily Kuang, Ehsan Jso. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. <https://www.microsoft.com/en-ww/microsoft-365/excel>.
- [53] Michael J. Kahn and Amanda Prail. *Formal Usability Inspections*, page 141–171. John Wiley & Sons, Inc., USA, 1994.
- [54] Takashi Kato. What “question-asking protocols” can say about the user interface. *International journal of man-machine studies*, 25(6):659–673, 1986.
- [55] Brandy Klug. An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience*, 1(6), 2017.
- [56] Ugur Kuter and Cemal Yilmaz. Survey methods: Questionnaires and interviews. *Choosing Human-Computer Interaction (HCI) Appropriate Research Methods*, 2001.
- [57] Clayton Lewis. *Using the” thinking-aloud” method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, 1982.
- [58] Clayton Lewis. *Using the” thinking-aloud” method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, 1982.
- [59] James R. Lewis. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.
- [60] James R Lewis and Jeff Sauro. The factor structure of the system usability scale. In *International conference on human centered design*, pages 94–103. Springer, 2009.
- [61] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

- [62] lookback. Lookback: simple, powerful, user research. <https://www.lookback.com/>.
- [63] Loop11. Online usability testing. <https://www.loop11.com/online-usability-testing-to-help-ux-marketing/>.
- [64] Johannes Manner, Stefan Kolb, and Guido Wirtz. Troubleshooting serverless functions: a combined monitoring and debugging approach. *SICS Software-Intensive Cyber-Physical Systems*, 34(2):99–104, 2019.
- [65] Bella. Martin. *Universal methods of design : 100 ways to research complex problems, develop innovative ideas, and design effective solutions / Bella Martin, Bruce Hanington*. Rockport Publishers, Beverly, MA, 2012 - 2012.
- [66] Maze. Maze — Product Research Platform for Modern Teams. <https://maze.co/>.
- [67] Jan Michels, Keith Hare, Krishna Kulkarni, Calisto Zuzarte, Zhen Hua Liu, Beda Hammerschmidt, and Fred Zemke. The new and improved sql: 2016 standard. *ACM SIGMOD Record*, 47(2):51–60, 2018.
- [68] Microsoft. Excel Spreadsheet Software. <https://github.com/WatVis/CoUX>.
- [69] Microsoft. Microsoft OneDrive - Personal Cloud Storage. <https://www.microsoft.com/en-gb/microsoft-365/onedrive/online-cloud-storage>.
- [70] Microsoft. Microsoft Sharepoint - Share files, build intranets. <https://www.microsoft.com/en-gb/microsoft-365/sharepoint/collaboration>.
- [71] Microsoft. .NET: Free. Cross-platform. Open source. A developer platform for building all your apps. <https://dotnet.microsoft.com/en-us/>.
- [72] Mozilla. JavaScript MediaRecorder API. <https://developer.mozilla.org/en-US/docs/Web/API/MediaRecorder>.
- [73] Michael D Myers. *Qualitative research in business and management*. Sage, 2019.

- [74] Walter Takashi Nakamura, Iftekhar Ahmed, David Redmiles, Edson Oliveira, David Fernandes, Elaine H. T de Oliveira, and Tayana Conte. Are ux evaluation methods providing the same big picture? *Sensors (Basel, Switzerland)*, 21(10):3480–, 2021.
- [75] Netlify. Develop and deploy web applications. <https://www.netlify.com/>.
- [76] NextAuth.js. Authentication for Next.js. <https://next-auth.js.org/>.
- [77] NextJS. The React Framework for Production. <https://nextjs.org/>.
- [78] Jakob Nielsen. *Usability Engineering*. Interactive Technologies. Elsevier Science & Technology, San Francisco, 1994.
- [79] Jakob Nielsen. Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414, 1994.
- [80] Jakob Nielsen. Ten usability heuristics, 2005.
- [81] Lene Nielsen, Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. Think-aloud surveys. In *IFIP Conference on Human-Computer Interaction*, pages 504–508. Springer, 2021.
- [82] Bevan Nigel, Kirakowskib Jurek, and Maissela Jonathan. What is usability. In *4th International Conference on HCI*, pages 1–6, 1991.
- [83] NodeJS. Node.js® is a JavaScript runtime built on Chrome’s V8 JavaScript engine. <https://nodejs.org/>.
- [84] Nodemailer. Email engine for Node.js. <https://nodemailer.com/>.
- [85] Mie Nørgaard and Kasper Hornbæk. What do usability evaluators do in practice? an explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 209–218, 2006.
- [86] npmjs. Node package manager. <https://www.npmjs.com/>.
- [87] NVivo. Best qualitative data analysis software for researchers. <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>.

- [88] Optimal Workshop. User Research Platform. <https://www.optimalworkshop.com/>.
- [89] José-Luis Padilla and Jacqueline P Leighton. Cognitive interviewing and think aloud methods. In *Understanding and investigating response processes in validation research*, pages 211–228. Springer, 2017.
- [90] PassportJS. Simple, unobtrusive authentication for Node.js. <https://www.passportjs.org/>.
- [91] Freddy Paz and José Antonio Pow-Sang. Current trends in usability evaluation methods: a systematic review. In *2014 7th International Conference on Advanced Software Engineering and Its Applications*, pages 11–15. IEEE, 2014.
- [92] S Camille Peres, Tri Pham, and Ronald Phillips. Validation of the system usability scale (sus) sus in the wild. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 192–196. SAGE Publications Sage CA: Los Angeles, CA, 2013.
- [93] PostgreSQL. The World’s Most Advanced Open Source Relational Database. <https://www.postgresql.org/>.
- [94] R.S. Pressman. *Software Engineering: A Practitioner’s Approach*. McGraw-Hill series in computer science. McGraw Hill, 2001.
- [95] Prisma Data Platform. Optimize Prisma for production workflows. <https://www.prisma.io/data-platform>.
- [96] Python. Python is a programming language that lets you work quickly and integrate systems more effectively. <https://www.python.org/>.
- [97] Qualtrics. Qualtrics - The leading experience management software. <https://www.qualtrics.com/>.
- [98] React. A JavaScript library for building user interfaces. <https://reactjs.org/>.
- [99] Sirpa Riihiäho. Usability testing. *The Wiley handbook of human computer interaction*, 1:255–275, 2018.

- [100] André Rodrigues, André Santos, Kyle Montague, Hugo Nicolau, and Tiago Guerreiro. Understanding the authoring and playthrough of nonvisual smartphone tutorials. In *IFIP Conference on Human-Computer Interaction*, pages 42–62. Springer, 2019.
- [101] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [102] Daniel Saldarriaga. Towards a tool to support think aloud and question-asking protocol evaluation with users. Informatics Project Proposal, School of Informatics, The University of Edinburgh, 2022.
- [103] Jeff Sauro and James R. Lewis. When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2215–2224, New York, NY, USA, 2011. Association for Computing Machinery.
- [104] Gonny LM Schellings, Bernadette HAM van Hout-Wolters, Marcel VJ Veenman, and Joost Meijer. Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European journal of psychology of education*, 28(3):963–990, 2013.
- [105] SmartLook. Session recording & behaviour analytics. <https://www.smartlook.com/>.
- [106] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. Coux: Collaborative visual analysis of think-aloud usability test videos for digital interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):643–653, 2021.
- [107] Andrew A Tawfik, Jessica Gatewood, Jaclyn J Gish-Lieberman, and Andrew J Hampton. Toward a definition of learning experience design. *Technology, Knowledge and Learning*, 27(1):309–334, 2022.
- [108] Teams. Make amazing things happen together at home, work, and school. <https://www.microsoft.com/en-gb/microsoft-teams/group-chat-software>.
- [109] The University of Edinburgh. DiscoverED. <https://discovered.ed.ac.uk/>.

- [110] Kelli Thoele, Mengmeng Yu, Mandeep Dhillon, Robert Skipworth Comer, Hannah L Maxey, Robin Newhouse, and Ukamaka M Oruche. Development and assessment of the usability of a web-based referral to treatment tool for persons with substance use disorders. *BMC Medical Informatics and Decision Making*, 21(1):1–12, 2021.
- [111] TryMyUi. Website usability testing — User Testing by TryMyUi. <https://www.trymyui.com/>.
- [112] TypeGraphQL. Prisma generator to emit TypeGraphQL type classes and CRUD resolvers from your Prisma schema. <https://prisma.typegraphql.com/>.
- [113] Usability Hub. User Research and Usability testing Platform. <https://usabilityhub.com/>.
- [114] User Feel. The better user testing tool. <https://www.userfeel.com/>.
- [115] User Zoom. Usability testing software for ux. <https://www.userzoom.com/>.
- [116] UserTesting. UserTesting — Hear what your audience is saying and see what they mean. So you can create better experiences. <https://www.usertesting.com/>.
- [117] UxCam. Deliver the perfect app experience. <https://uxcam.com/>.
- [118] UXTweak. Powerful tools for UX research & user testing — UXTweak. <https://www.uxtweak.com/>.
- [119] M.J. van den Haak and M.D.T. de Jong. Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. In *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.*, pages 3 pp.–, 2003.
- [120] Lex van Velsen, Thea van der Geest, and Rob Klaassen. Testing the usability of a personalized system: comparing the use of interviews, questionnaires and thinking-aloud. In *2007 IEEE International Professional Communication Conference*, pages 1–8. IEEE, 2007.
- [121] Vercel. Next.JS Commerce: The all-in-one React starter kit for high-performance ecommerce sites. <https://nextjs.org/commerce>.

- [122] Vercel. Vercel combines the best developer experience with an obsessive focus on end-user performance. <https://vercel.com/>.
- [123] Suzanne E Wade. Using think alouds to assess comprehension. *The Reading Teacher*, 43(7):442–451, 1990.
- [124] Cathleen Wharton, John Rieman, Clayton Lewis, and Peter Polson. The cognitive walkthrough method: A practitioner’s guide. In *Usability inspection methods*, pages 105–140. 1994.
- [125] Michael Winter, Harald Baumeister, Ulrich Frick, Miles Tallon, Manfred Reichert, and Rüdiger Pryss. Exploring the usability of the german covid-19 contact tracing app in a combined eye tracking and retrospective think aloud study. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2215–2221. IEEE, 2021.
- [126] www.10up.com. System Usability Scale - Score. <https://10up.com/uploads/2018/11/sus-score-1-768x427.jpg>.
- [127] Yi Zhang. The design of a mobile app to promote young people’s digital financial literacy. In *International Conference on Human-Computer Interaction*, pages 118–136. Springer, 2021.
- [128] Tingting Zhao and Sharon McDonald. 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*, pages 581–590, 2010.
- [129] Zoom. Video conferencing. <https://zoom.us/>.

Appendix A

Background and related work

A.1 Usability inspection methods

- Heuristic evaluations [79] is a method in which an expert in usability reviews each part of an interface and assesses if that part of the interface complies with a set of usability criteria, commonly known as best practices. Usually, those good practices involve Nielsen's ten heuristics [80].
- Cognitive walkthroughs [124] are a method in which an expert in usability takes the role of a potential user of a particular system. Usually, the expert executes specific tasks that the user would typically execute. During that process, experts find issues in the system's usability and provide feedback to the designers or developers.
- Feature inspections [79] are similar to cognitive walkthroughs, but instead of executing specific tasks, the expert inspects the system for specific features. Usually, the expert checks the inputs required by the feature and compares the output given by the system with the expected output and draws conclusions based on the results.
- Pluralistic walkthroughs [12] do not involve a single person evaluating the system. Instead, it considers a team of different experts, designers and developers who gather together to execute tasks in the system, allowing them to reach a consensus on which features should be improved and which usability improvements can be made.

A.2 Usability evaluation tools

A.2.1 Evaluated usability testing tools

The first of those tools evaluated is called Maze [66]. It is a tool aimed to test Figma [33] or AdobeXD [2] prototypes. Users can create new *Mazes*, the concept they use for defining a usability study. Within a *Maze*, a user can input the link of the design prototype, can set up tasks for a participant and can create a questionnaire for the end of the evaluation. In Maze it is possible to send a link to several participants, and they offer a paid *panel* of users, which researchers can leverage for obtaining more feedback. Maze also offers a feature to analyse the results, but, at least on the free trial, it is not possible to export the data.

The second tool evaluated is called UserTesting [116]. They do not offer a free trial, so all the information gathered was obtained directly from their website. In UserTesting, a researcher can input a website, app or prototype, and set up different tasks for users to execute. UserTesting provides demographic information of the participants and gives screen, audio and video feedback of what the participants did when using the product. UserTesting claim to have a *global network of contributors*, meaning that researchers can specify which type of users they want, and UserTesting will provide participants with those characteristics. In terms of data analysis, UserTesting offers automatic transcripts and sentiment analysis. Their website is not explicit in terms of the pricing, but they claim to have flexible pricing in terms of the required testing capacity.

The third tool evaluated is called UxTweak [118]. It is the only tool from the five tools evaluated that explicitly allows Think Aloud [57]. Similar to the tools explained before, it is possible to set up a website on UxTweak and define the tasks that the participant needs to execute. For evaluating a website, developers must install UxTweak's plug in in the code, so it is a tool that requires the involvement of technical teams. UxTweak also has a pool of participants, that researchers can pay for to access. In UxTweak it is possible to select different kinds of usability evaluation methods, including Think Aloud [57], as explained before. When a participant finishes a Think Aloud [57] session, UxTweak provides audio and video recordings of the participants' interactions and also provides a way of exporting data. It is important to note that the data export feature is not available on the free trial version of the platform.

The fourth tool that was reviewed is called TryMyUi [111]. It is a tool specifically designed for testing websites, so it does not allow apps nor prototypes. Similar to the tools described above, it is possible to define tasks and to pay for recruiting participants.

In terms of data analysis, TryMyUi is the only reviewed tool that leverages the System Usability Scale [16] and presents it as a result. It also allows data export and automatic transcriptions but not on the free trial.

Finally, I also checked Lookback [62], a very similar tool that also allows website testing. Similar to TryMyUi [111] and UxTweak [118], Lookback [62] also requires a script to be installed on the analysed website. Similar to UserTesting [116] and unlike the other tools reviewed before, Lookback offers both moderated and unmoderated tests, which means that an expert can be present or not in the session with the participant. In terms of data analysis, Lookback does not offer data export functionalities, at least on the free trial version, but researchers can see insights when participants finish their tasks.

A.2.2 Additional usability evaluation tools

1. Optimal Workshop: allows different methods such as Card sorting, surveys and first-click testing. Also has a user pool that experts can buy [88].
2. Usability Hub: allows surveys, five-second tests, fist-click tests and prototype user testing [113].
3. Loop11: focused on testing wireframes, prototypes and website initial iterations. Also features a user pool that experts can pay for [63].
4. Userfeel: offers a pay-per-evaluate method in which a researcher can pay for accessing not only a participant pool but also an expert pool, which will use different methods to assess the usability of a system [114].
5. Hotjar: provides a way to assess the usability of a website by checking users' heatmaps and screen recordings. Provides also an interface for surveying users [48].
6. UserZoom: provides different methods for doing usability evaluation like card sorting, surveys and *generic usability testing* with tasks and video recording [115].
7. CrazyEgg: features heatmaps, recordings, A/B testing and surveys for analysing the usability of a website [23].

8. FiveSecondTest: leverages the five second test method and offers insights on user interaction using that method [35].
9. UxCam: specialises on analytics obtained from heatmaps and video recordings from user sessions. Also offers event analytics [117].
10. Smartlook: similar to UxCam, also offers data based on session recordings, events and heatmaps [105].

Appendix B

Requirement gathering and design

B.1 Initial sketch of Wanda

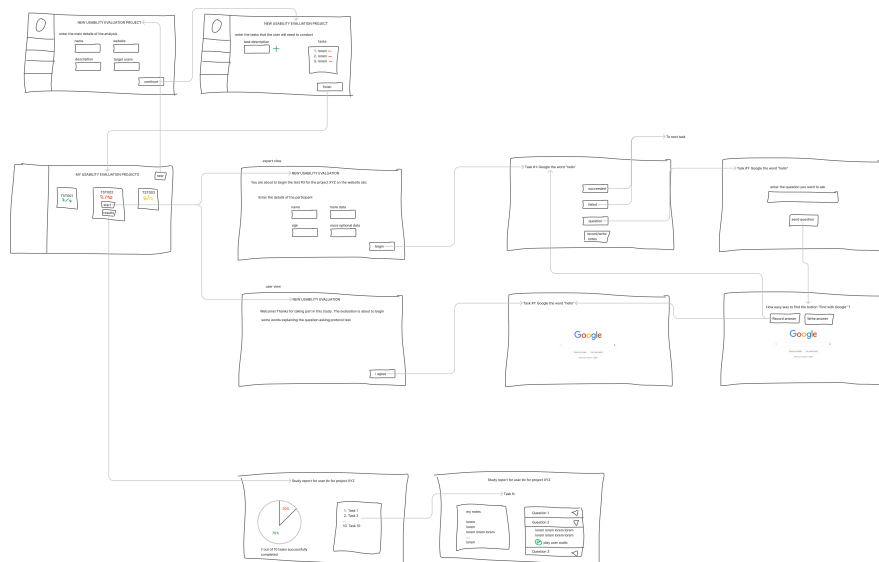


Figure B.1: Initial sketches of Wanda. Self drafted using Figma [33]

B.2 Participant recruitment

We recruited 5 HCI experts from The University of Edinburgh with vast experience in usability. They had all conducted usability evaluations before either in academic or industrial environments, and they all had experience with either Think Aloud [57] or Question-Asking Protocol [54].

B.3 Data collection method

As explained before in section 2.2, I conducted one semi-structured interview with each HCI expert. All the sessions were online, using Teams [108]. The interviews took 30 minutes, and the sessions were recorded and stored on the University's Microsoft SharePoint [70] server.

B.4 Materials

For this study, I built a Participant Information Sheet that can be found on the G.1 and a Participant Consent Form that can be found in the appendix H.1. Since the interviews were semi-structured, I set up a list of questions that was used as a base for guiding the conversation, but with each expert, the order of the questions was changed to ensure that the flow of the interview was maintained. The base list of questions can be found on the following section.

B.4.1 List of questions

Semi-structured interview questions

- General description of the project and socialization of the meeting objective.
 - Consent to record the meeting.
1. Please describe the general process and stages you follow to evaluate the usability of a digital tool.
 2. Have you ever used think-aloud and/or question-asking protocol? If so, how was your experience?

Specifically for the think-aloud method:

1. What are the main advantages of using this method?
2. What are the main disadvantages of using the method?
3. How do you design and prepare the experiment? Do you use any particular tool?
4. How do you analyse the data collected during the tests? Do you use any specific tool?
5. What is the main challenge of processing usability test data using the think-aloud method?

Specifically for the Question-Asking protocol

1. What are the main advantages of using this method?
2. What are the main disadvantages of using the method?
3. How do you design and prepare the experiment? Do you use any particular tool?
4. How do you process the data collected during the tests? Do you use any specific tool?
5. What is the main challenge of processing usability test data using the Question-Asking Protocol?

Tool features

1. What features would you like to have?
2. What features do you think are not necessary?
3. What recommendations do you have for improving the usability of our tool?
4. How the evaluation preparation should look like
5. How data analysis should look like

B.5 Procedure

First, all the experts were contacted by e-mail to check their availability and desire to help with the project. Once they agreed to participate in the study, I sent them the Participant Consent Form, the Participant Information Sheet, a time schedule, and the Teams [108] link to the session.

Once the sessions began, I would introduce the objective of the meeting and the objective of the project. Before I began to ask the questions, I made sure that they filled the Participant Consent Form and that they agreed to being recorded. After that I would turn on the recording, and begin by asking them about their experience with Think Aloud [57] and Question-Asking Protocol [54].

After they told me about their experience, I would go further down the questions depending on their previous answers, trying to get more detail about the specific methods, tools and methodologies they used on each type of evaluation they had experience with.

In the end, I asked them about potential features for the tool, starting from features for the evaluation design, continuing with features for the execution of the evaluations and finishing with suggestions they might have for data analysis within the tool.

After each evaluation, I moved the video recordings from Microsoft OneDrive [69] to Microsoft SharePoint [70], to avoid the risk of data loss due to OneDrive's retention policy of two months at the time of using the software. Once the videos were safely stored, I would extract the transcript of the videos to be able to analyse them faster.

B.6 Data analysis

In the figure ?? I present the code map that I used for analysing the data of the requirement gathering phase using Thematic Analysis [19].

Each of the themes identified had different frequencies throughout the evaluation. In the figure B.8 I present the hierarchy of the themes identified. It can be seen from the chart that the experts widely commented on the features for the tool and that they had also broad examples on how to prepare before the evaluation. It can also be inferred from the figure that experts were not particularly interested in commenting about Question-Asking Protocol [54], which could mean that some of them don't have vast experience with the method.

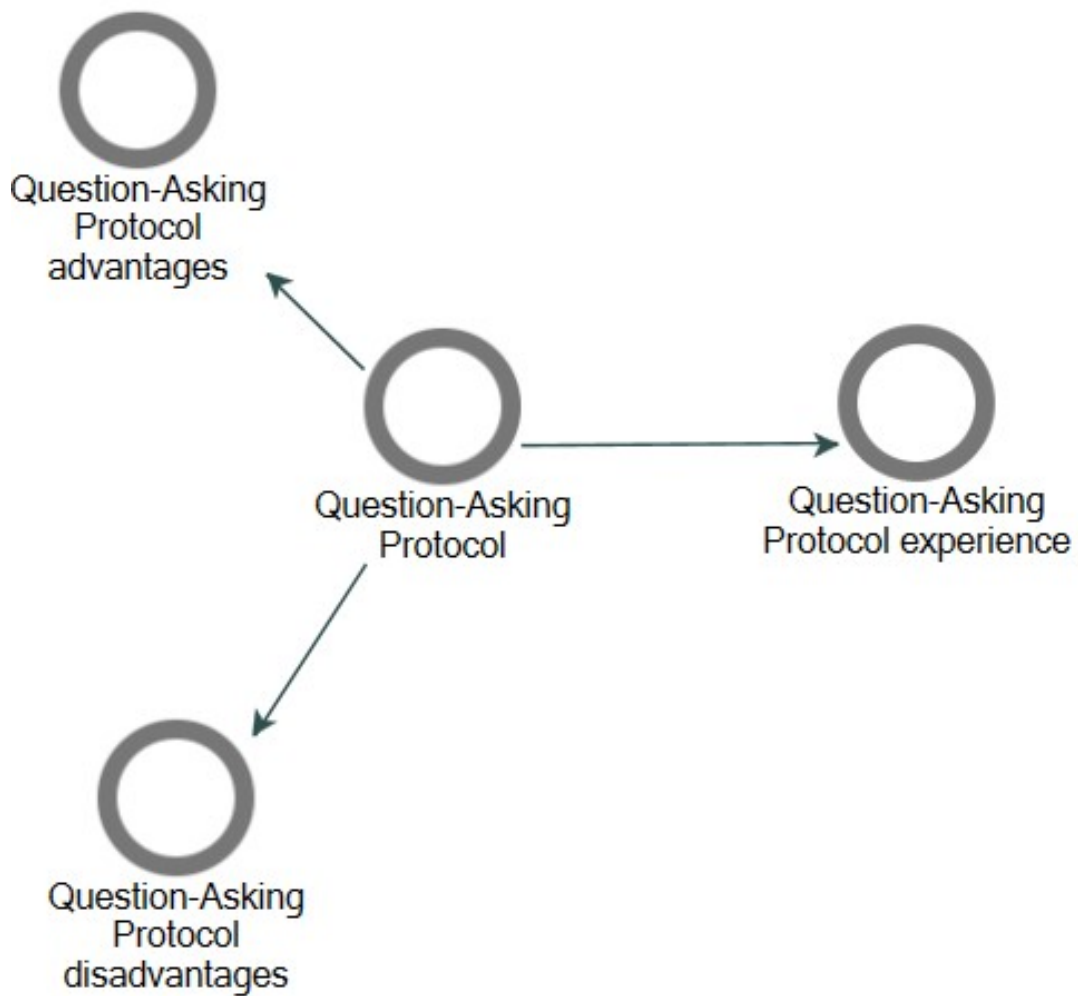


Figure B.2: Question-Asking Protocol [54] themes. Built using NVivo [87]

B.7 Results

In general, all the experts were very open with their experience and were actively willing to provide their opinion on what the tool should look like. All the experts had conducted Think Aloud [57] sessions before, but only one had conducted Question-Asking Protocol [54] sessions.

B.7.1 Think Aloud and Question-Asking Protocol methodology

B.7.1.1 Experience with the methods

Two of the five experts had used Think Aloud [57] intensively before; two said they had only used it in academic environments, and one said they used it once but did not like

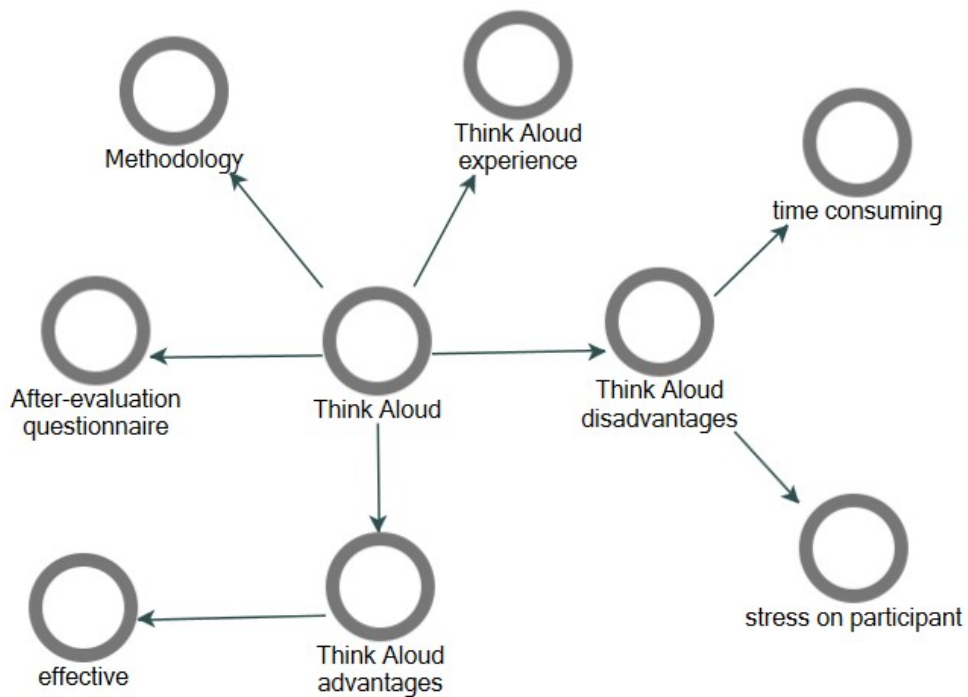


Figure B.3: Think Aloud [57] themes. Built using NVivo [87]

the method. Of the experts that expressed vast experience with the method, the most common applications were educational or telehealthcare systems, where the method would be used to gather feedback on the early stages of design.

In academic environments, the experts that claimed to have used the Think Aloud [57] method said that it could be used in conjunction with other methods or technologies such as eye-tracking or speech-to-text software. Also, two experts said they had used the method while teaching students how to use it in classroom environments.

Only one expert manifested experience with Question-Asking Protocol, and they said only in academic environments where students used the method for diverse projects.

“I used think aloud [at] different stages in the design of my PhD in 2010 [...] and then I used it in various other projects. But I mostly use [think aloud] in the formative evaluation stage after designing the low fidelity prototype.” - E3.

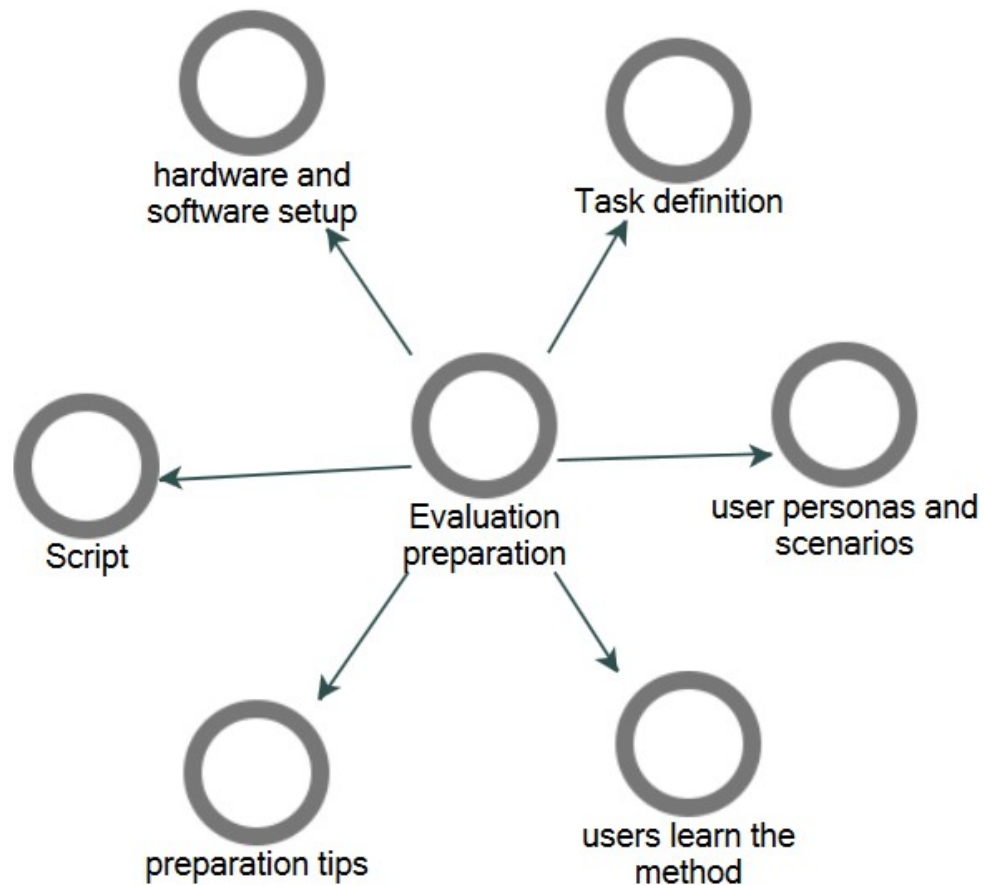


Figure B.4: Evaluation preparation themes. Built using NVivo [87]

B.7.1.2 Advantages and disadvantages of Think Aloud and Question-Asking Protocol

All but one expert manifested advantages of the Think Aloud [57] method. The most common advantage expressed by the experts is that Think Aloud [57] is an efficient method for gathering user feedback. Two experts said it is a relatively quick method and easy to use. One expert said that since it is an uninterrupted stream of thought, evaluators can get much information that otherwise would not be present.

As stated before, one expert said they dislike the method for two main reasons that other experts supported. On the one hand, four experts expressed that Think Aloud [57] can strain the participant because they might feel observed or examined, and many participants could find that situation very uncomfortable. On the other hand, three experts expressed that the data produced by a Think Aloud [57] session is very tedious to process and analyse.

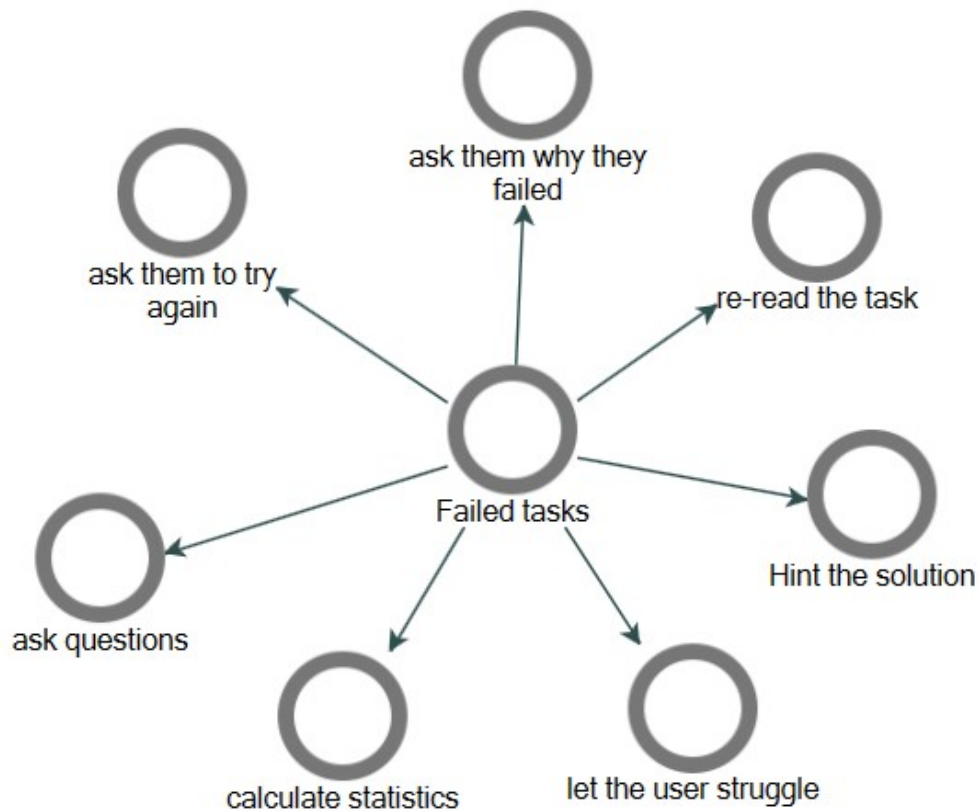


Figure B.5: Failed task management themes. Built using NVivo [87]

The expert with experience in Question-Asking Protocol expressed that they usually prefer that method over Think Aloud [57] because it allows examiners to guide the session through questions to find the most relevant feedback from participants. Since on Think Aloud [57] examiners are not supposed to intervene until the end of the session, some interesting observations may be left out of the analysis, potentially losing some valuable information.

“Some people are somehow reluctant to express their thoughts because they think that they can be judged.” - E2.

B.7.1.3 Study preparation

When asked how they prepare for the studies, they all agreed that the most crucial part is to do a well-thought definition of the tasks. They explained that the tasks need to be prepared depending on the primary goal of the system or the main features that need to be evaluated, but they also expressed that preparing the tasks needs attention and must

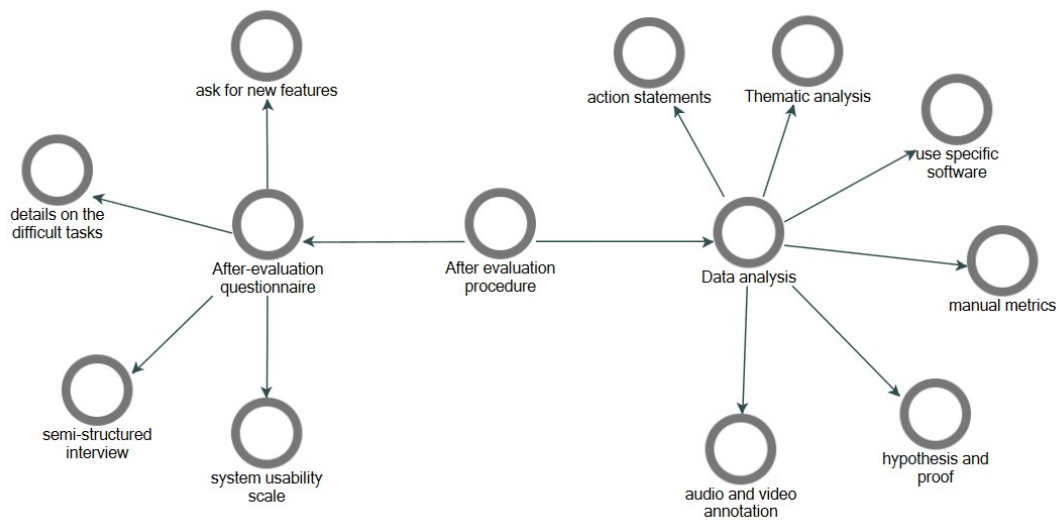


Figure B.6: Questionnaires and data analysis themes. Built using NVivo [87]

be a careful process. Two experts suggested that the tasks should be pointed towards the most executed tasks in the system, while two suggested that the tasks should be pointed towards the essential features of the system, which are not necessarily the same thing.

The experts advised on how to define the tasks. They mostly agree that the redaction of the task should be goal-oriented instead of detail-oriented. They also pointed out that the tasks should be clear enough for participants without prior experience with the system. Additionally, one expert explained that the duration of the tasks is also crucial, so they recommended doing a pilot before the evaluation to make sure that the task flows and durations make sense.

Three experts also said that the script for the evaluation is vital in letting the user know how to conduct the evaluation. The script, as they explained, is a document that is shown to the participant before the session starts. It is crucial to have well-defined scripts because they allow participants to get familiar with the method and reduce potential stress on the participant.

Finally, one expert said that it is also essential to define the research question and aim the tasks toward answering that question. They expressed that the research question gives the examiner a clear view of the objective of the evaluation and that having the objective explicit allows defining the tasks to ensure that they answer the research goal.

“You basically define tasks that correspond with the main features that you have developed in your system, making sure that the tasks do not mention specific interface elements like buttons or scrolling or menus [...] so that they’re kept quite high level and

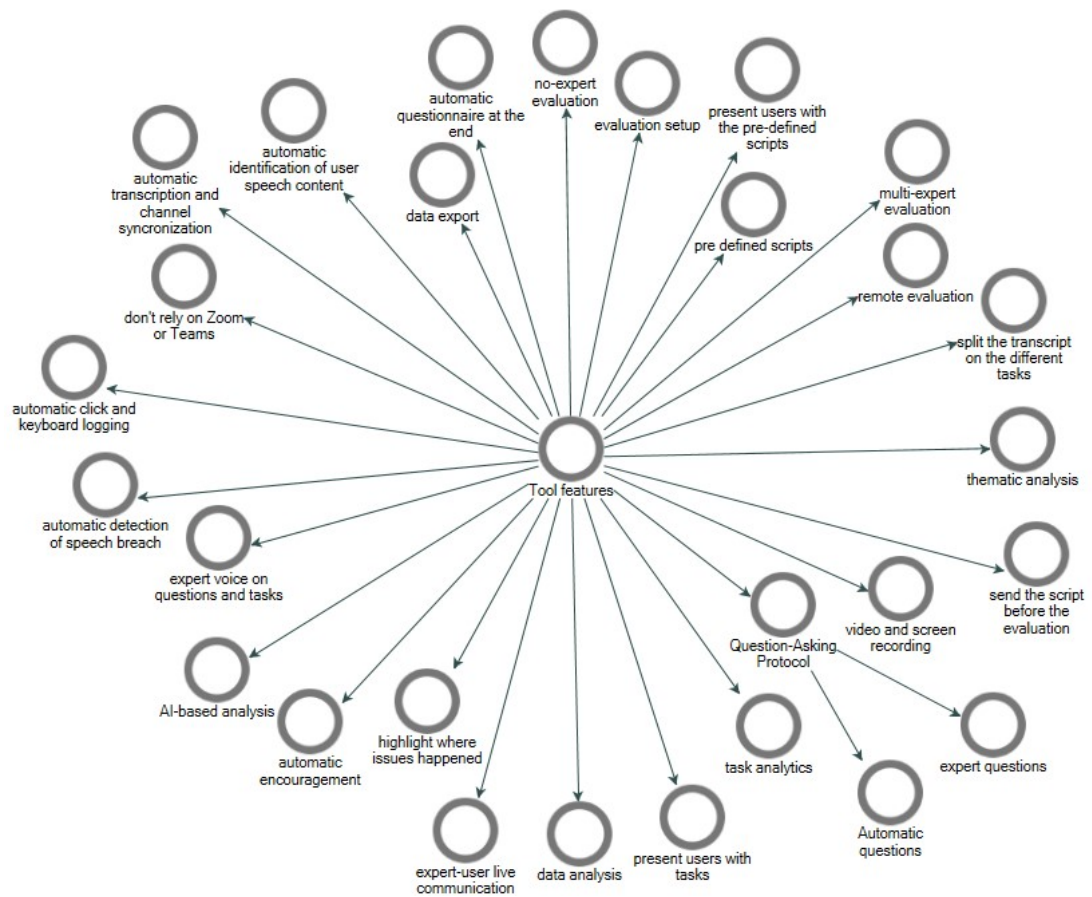


Figure B.7: Tool-related themes. Built using NVivo [87]

abstract so that they figure out the specific things themselves” - E1.

B.7.1.4 Failed tasks

When asked how they handle the situation when a user fails the task, the most common answer was to let the user struggle for some time. They expressed that it is beneficial to wait until the participant finds for themselves how to get the task done because that way, the examiners can often identify what the underlying usability issue that is causing the user to struggle is.

Some experts suggested that it is possible to give hints to the participants, but they were divided regarding whether or not it is beneficial. Two experts said hinting to the user on how to finish the task may aid in the agility of the evaluation and may make the participant more comfortable with the task. However, two experts also pointed out that giving hints should be an extreme case or the last resource because, as an evaluator, you may lose valuable information on how the participant interacts with the interface.

When a participant fails the task, two experts suggested asking participants questions

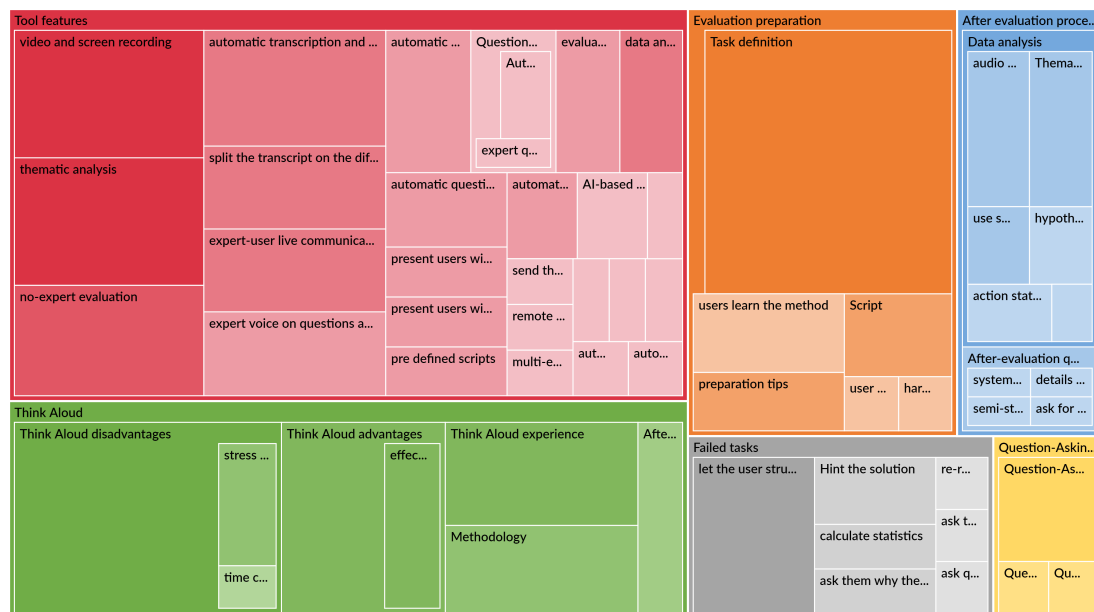


Figure B.8: Requirement gathering theme hierarchy

to gather feedback on why they think they failed. Asking questions may be beneficial for understanding the root cause of the failure and ultimately identifying the underlying usability issue.

“I think that you probably want to get the worst case scenario. [...] and in that case you want to leave them alone for quite a long time until you know until it’s clear that they really can succeed” - E4.

B.7.1.5 After-evaluation Procedure

Experts agreed that after a Think Aloud [57] session, it is recommended to ask a questions to the participant. It is frequently a pre-defined list of questions that is designed to get data about the main features of the system, but all experts suggested that it is common to add questions to the list during the session. It is usually a semi-structured interview in which the participant can reflect on additional thoughts of the system they have just evaluated.

One expert also suggested that it is customary to add the standard questionnaire of the System Usability Scale [16], since it can be useful to obtain a quick insight on the general usability status of the evaluated system. Additionally, one expert suggested that it is also useful to ask questions about specific features where there are implementation doubts.

“You could ask specific questions [...] about particular features that [you] weren’t sure how to implement or where [you] had several options, and getting the users to give their opinion on them specifically” - E1.

B.7.1.6 Data Analysis

After both the evaluation and the semi-structured interview have finished, the examiners need to conduct data analysis. For doing so, the most common method is to perform Thematic Analysis [19], which requires the examiners to be able to transcribe all the audio material from the session. Thematic Analysis, however, was discouraged by one expert arguing that it may be unnecessary since experts already have a lot of information about the reasons for users’ reactions.

“It’s the sort of situation where it’s not like a social science experiment or something like this. You don’t necessarily want to get into some of the usual processes of being with qualitative data like doing thematic analysis [...], These kinds of things might be unnecessary in the sense that you have quite a lot of knowledge about what the the situation is and what and why you get the the particular sorts of reactions that you do.” - E2.

Two experts expressed having used specific software for processing data. Amongst the descriptions they gave, they usually use software that automatically allows them to transcribe the audio files, but they expressed that those softwares are not specifically designed for Think Aloud [57] and that they may be sometimes expensive.

“I normally you would transcribe the whole audio [...] and then I would do thematic analysis.” - E1.

B.7.2 Feedback about tool features

The questions about the tool features were divided into three sections. First, the experts were queried about features for the design of an evaluation study. For this case, in addition to preparing the tasks, two experts suggested having pre-defined scripts, one expert suggested double-checking the difficulty of the tasks, and one expert expressed that it would be good not to rely on Zoom [129] or Teams [108] for actually conducting the evaluation.

Regarding the evaluation execution with the participants, all experts agreed that the most important thing is to keep an audio and video recording of the session for analysing

data afterwards. Two experts mentioned it would be good to allow the participant to hear the instructions with the expert's voice instead of having to read them. Two experts said that it would be good to do automatic click and keyboard logging. Two experts expressed interest in having automatic encouragement, and one mentioned that it would be good to detect speech breaches automatically.

Experts also gave feedback on the post-evaluation procedure. All of them suggested that after a Think Aloud [57] session, it is advised to include a questionnaire, which should be prepared beforehand. However, three experts said that it would be good to be able to add questions to the questionnaire during the evaluation in case they notice something they want to ask afterwards. Regarding data analysis, all the experts said that it is crucial to have automatic transcripts of the participants' verbalisations. Three experts recommended splitting the transcripts by task to ease the analysis process. Regarding analysis methods, four experts commented on thematic analysis but had divided opinions. Two experts said they would use thematic analysis, but two explicitly said they would not.

Finally, experts were asked about the possibility of allowing participants to execute evaluations independently without needing an expert. This question comes from the market benchmark since some tools like UxTweak [118] allow such features. Four experts commented on the topic, and they also had divided opinions. Two experts were in favour of implementing such a feature. One expert said it would be good but alerted that it is a trade-off between the amount of data you get versus the quality of the data. Finally, one expert commented that the feature would be interesting, but they suggested it is out of the scope of an MSc dissertation project.

“Just the participants themselves? You know, yeah, maybe. Think about this: distributed participatory design when people are all over the the world, interesting.” - E3

B.7.3 User Stories

From the given by the experts, a list of 27 user stories was built and prioritised. From those user stories, five were discarded due to the technical difficulty considering the timeframe of the project. The full list of final user stories can be found on appendix B.8

B.8 User Stories

1. As an HCI expert, I want to setup a new remote evaluation so that I can reach more users for my study.
 - (a) As an HCI expert, I want to define the tasks that the users would need to follow in the study, so that I can evaluate the most important features in my system.
 - (b) As an HCI expert, I want to setup a name for my study, and also which website I want to evaluate, so that I can identify where can its usability be improved. I also want to input the target number of participants.
 - (c) As an HCI expert, I want to define the goal or research question of my study, so that I can draw conclusions at the end.
 - (d) As an HCI expert, I want to choose a pre-defined script to show to the users of my study, so that they can get familiarized with the evaluation method.
 - (e) As an HCI expert, I want to set up a questionnaire for presenting the user at the end of the evaluation, so that they can give me more detailed feedback.
2. As an HCI expert, I want to conduct a remote evaluation of my website with a user, so that they can test the website and I can identify usability improvements.
 - (a) As an HCI expert, I want to send a link to my user so that they can connect to the evaluation session.
 - (b) As a user, I want to define when to begin the session, so that I am sure that I understood the method.
 - (c) As a user, I want to play a recording of the script, so that I can hear it from the expert's voice.
 - (d) As a user, I want to see which task am I supposed to execute, so that I don't get lost in the objective of the evaluation.
 - (e) As a user, I want to listen to the task instead of reading it, so that it is easier for me to execute it.
 - (f) As a user, I want to be able to manually mark a task as finished, so that the evaluation continues with the subsequent task.
 - (g) As an HCI expert, I want to decide if the user's task was succeeded or failed after they mark it as finished, so that I can identify them later.

- (h) As a user, when the evaluation finishes, I want to see which question am I supposed be answering so that I can give the correct answers.
 - (i) As a user, I want to see a summary of all the questions I answered.
3. As an HCI expert, I want to send the user a link for performing a stand-alone evaluation without my presence, so that I can reach more users in less time
- (a) As a user, I want to connect to the session I received from the expert, so that I can conduct the evaluation.
 - (b) As a user, I want the system to automatically present me with the tasks, so that I can follow them in the correct order.
 - (c) As a user, I want to define a task as failed or succeeded after I finish it, so that an expert can later analyse them.
 - (d) As a user, I want the system to automatically show the questionnaire when all the tasks are done, so that I can complete it.
4. As an HCI expert, I want to analyse the data after the evaluation sessions
- (a) As an HCI expert, I want to read the transcript of the audio of the session, so that I can process it faster.
 - (b) As an HCI expert, I want to have the audio transcript splitted by each of the tasks of the evaluation, so that I can process them faster.
 - (c) As an HCI expert, I want to visualize charts with the most important results of the evaluation study.
 - (d) As an HCI expert, I want to see the System Usability Scale [16] result if I set it up at the beginning of the evaluation.
 - (e) As an HCI expert, I want to be able to export the data for a session and for the complete evaluation study.

B.9 Prioritized requirements

User story	Section	Priority	Reason for priority
Task definition	Evaluation preparation	High	Supported by experts 1 and 3
Input website	Evaluation preparation	High	Required by any usability evaluation system
Define the research question	Evaluation preparation	Medium	Supported by expert 5 with the argument that it makes it easier to define the tasks
Define evaluation script	Evaluation preparation	Medium	Supported by experts 1 and 4

User story	Section	Priority	Reason for priority
After-evaluation questionnaire	Evaluation preparation	High	Supported by experts 1, 4 and 5.
Send the evaluation link to participants	Evaluation execution	High	Required by any remote usability evaluation study
Live communication with the participant	Evaluation execution	Low	Supported by experts 2, 4 and 5, but may be very challenging in the time frame
User consent for starting session	Evaluation execution	Low	Supported by experts 2 and 3
Play a recording of the script	Evaluation execution	Medium	Supported by expert 1
See current task	Evaluation execution	High	Required by any remote usability evaluation study
Listen to the task description	Evaluation execution	Medium	Supported by expert 1 and 2. Useful for differentiating from other usability study platforms and improving ux.
Mark a task as finished (participant)	Evaluation execution	High	Required by the Think Aloud [58]method.
Mark a task as finished (expert)	Evaluation execution	High	Required by the Think Aloud [58]method.
Manually remind the user to speak aloud	Evaluation execution	Low	Supported by experts 1 and 3
Add questions to the questionnaire	Evaluation execution	High	Supported by all the experts.
See current question (participant)	Evaluation execution	High	Required to fill the questionnaire.

User story	Section	Priority	Reason for priority
No-expert evaluation	Evaluation execution	High	Supported by experts 1, 3, 4 and 5
Automatic identification of participant speech breach	Evaluation execution	Discarded	Supported by expert 1. Not feasible in the time frame of the project.
Don't rely on Zoom or Teams for live communication	Evaluation execution	Discarded	Supported by expert 1. Not feasible in the time frame of the project.
View video recording	Post-evaluation	High	Supported by all the experts.
Read audio transcript	Post-evaluation	High	Supported by all the experts.
Split audio transcript by task	Post-evaluation	High	Supported by all the experts.
Data export	Post-evaluation	Medium	Supported by expert 5.
Cross-participant analytics	Post-evaluation	Low	Supported by expert 5.
Thematic analysis	Post-evaluation	Discarded	Supported by experts 1 and 3, discouraged by experts 2 and 4.
AI-based analysis	Post-evaluation	Discarded	Supported by expert 4. Not feasible in the time frame of the project.
Automatic identification of user speech content	Post-evaluation	Discarded	Supported by expert 1. Not feasible in the time frame of the project.

Appendix C

First iteration of development

C.1 Back-end

C.1.1 E-mails

For sending e-mails, Wanda is relying on the package *nodemailer* [84] with an e-mail server configured in AWS's [9] SES service. The system sends e-mails in two situations: First, when a user logs in using the e-mail provider, the system sends an e-mail with the link to sign in. Second, when an expert invites a participant to a new study session, the system sends an e-mail to the participant with the link for connecting to the session. AWS SES was the correct choice for this project because it offers a free tier allowing 62.000 messages per month, which is more than enough for the initial phases.

C.2 Authentication

When users go to the landing page, they find a button that allows them to choose which method they want to use for logging in. For this purpose, I implemented three specific methods: Google [41], GitHub [38] and e-mail magic links, but NextAuth has more than 50 different authentication providers that can be used. All of them are based on the OAuth2 [44] protocol, which is an industry-wide standard for authentication.

Since Next.js [77] and NextAuth [76] are full-stack frameworks, the system can share authentication information in both the front-end and back-end. NextAuth exposes a hook called *useSession* to the front-end and a method called *getSession* to the back-end. Those methods are used within the system to check whether the user is authenticated or not. Additionally, those methods are also helpful for ensuring that the user has the

correct permissions to access the resources they intend to.

C.2.1 Table definition

- **EvaluationStudy**: holds the studies that experts create. Each study can be understood as a website that the expert wants to analyse.
- **Task**: holds the *tasks* that are included in each *Evaluation Study*. These are the tasks that participants will need to accomplish when they evaluate the that the expert set up.
- **Script**: holds *scripts* that can be used for giving instructions to the participants when they begin to evaluate a website.
- **Questionnaire**: holds the *questionnaires* that will appear to the participant at the end of the evaluation.
- **Question**: holds all of the *questions* that are part of a *questionnaire*. Also, defines if a question is part of the System Usability Scale [16] or not.
- **StudySession**: holds all the *sessions* in which participants can connect and conduct an evaluation on the website defined by the expert.
- **StudySessionTask**: holds all the *tasks* assigned to a participant within a *study session*.
- **QuestionResponse**: holds all the *responses* a participant has given on a specific *question* within the *questionnaire* of a *study session*.
- **StudySessionData**: holds transactional data about the *study session* of a participant.
- **User**: holds all the *users* that are part of the system. Both participants and experts.
- **Account**: holds all the *accounts* that a *user* can have. For instance, a user can log in using Google, GitHub or their e-mail, so each account is stored individually.
- **Session**: holds all the *sessions* a user has active. A session means an active log in that the user performed, so a user may have an active log in in many devices.
- **Role**: holds all the *roles* that the system includes. For instance, this table only has three rows: Administrator, Expert and Participant.

- RoleToUser: holds all the *roles* that a *user* has. For example, a user can be both an *expert* and a *participant* in a session of a fellow expert.
- Page: holds all the *pages* that are part of the website.
- PageToRole: holds all the *pages* that a *role* has access to.
- Materialized view EvaluationStudyResult: holds a materialized query of the results of a *study session* conducted by a participant.

C.2.2 Database ER Diagram

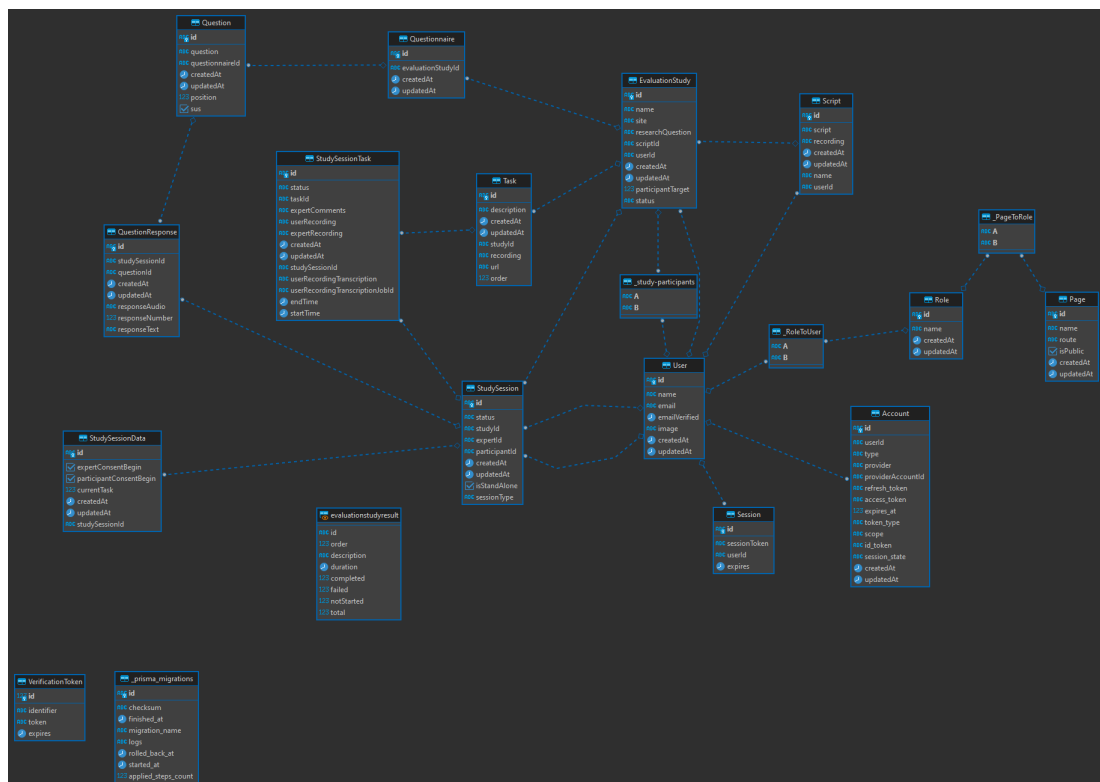


Figure C.1: Database ER Diagram

C.3 Front End

In this section I present the most relevant pictures of Wanda's interface.

C.3.1 Landing and Authentication

The following pictures correspond to Wanda's features that are common to both experts and participants.

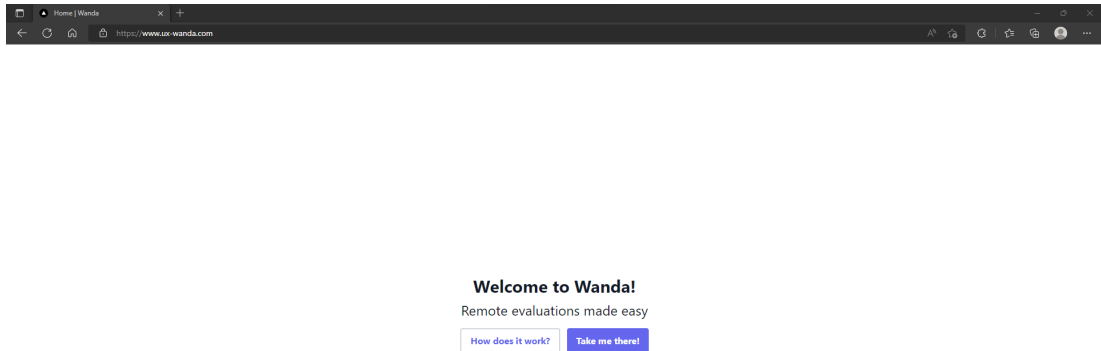


Figure C.2: Wanda landing page. First iteration of Wanda.

C.3.2 Participants' interface

The pictures shown below correspond to the participants' interface at the stage of the first iteration of Wanda.

C.3.3 Experts' interface

The pictures shown below correspond to the experts' interface at the stage of the first iteration of Wanda.

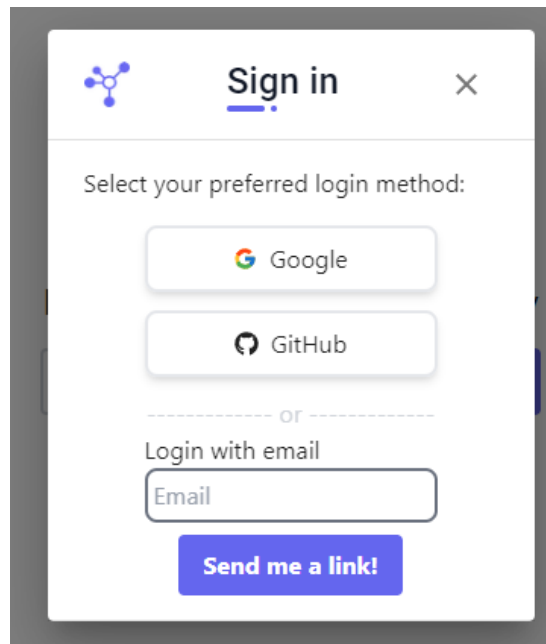


Figure C.3: User sign in form. First iteration of Wanda.

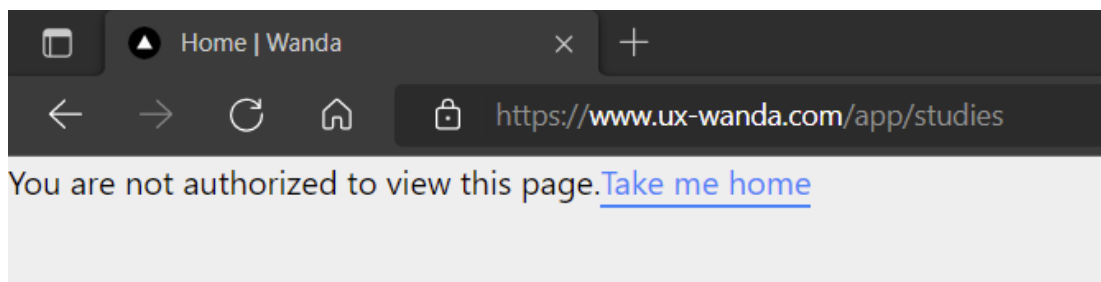


Figure C.4: Users arriving to a page where they don't have access. First iteration of Wanda.

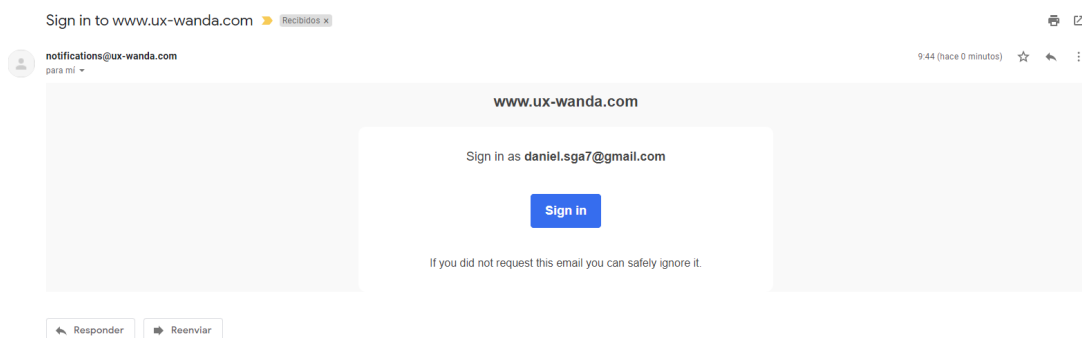


Figure C.5: Log in e-mail sent to the user with the magic link. First iteration of Wanda.

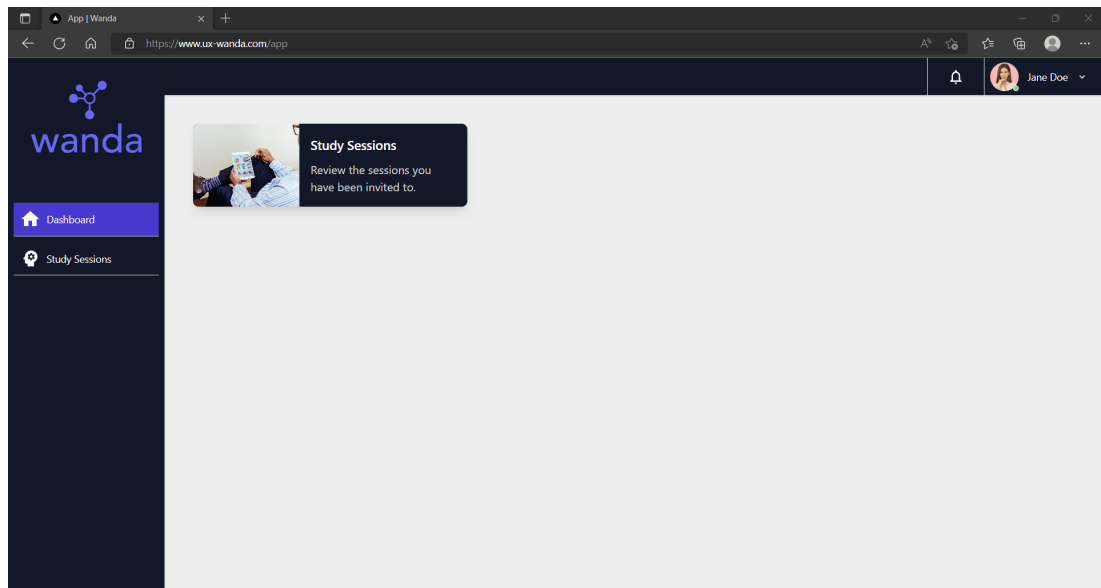


Figure C.6: Participants' homescreen. First iteration of Wanda.

Status	Evaluation study	Participant	Actions
STARTED	ComposerFM	danielsga7@gmail.com	🗑️
STARTED	E-commerce evaluation	danielsga7@gmail.com	🗑️
COMPLETED	E-commerce Evaluation	danielsga7@gmail.com	🗑️
COMPLETED	E-commerce Evaluation	danielsga7@gmail.com	🗑️

Figure C.7: Participants' study session table. First iteration of Wanda.

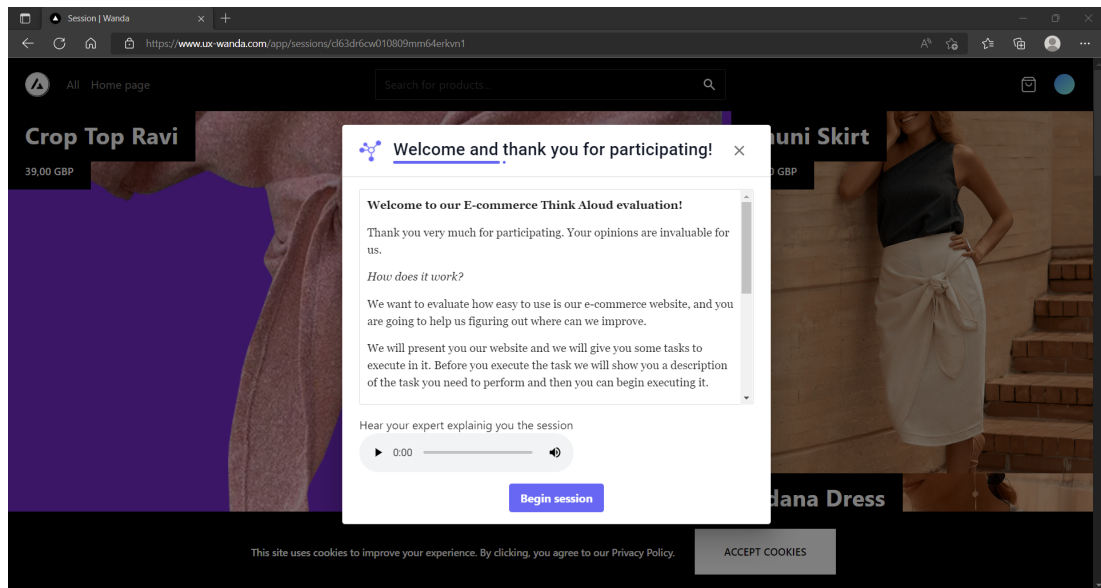


Figure C.8: Participants' study session script and begin action. First iteration of Wanda.

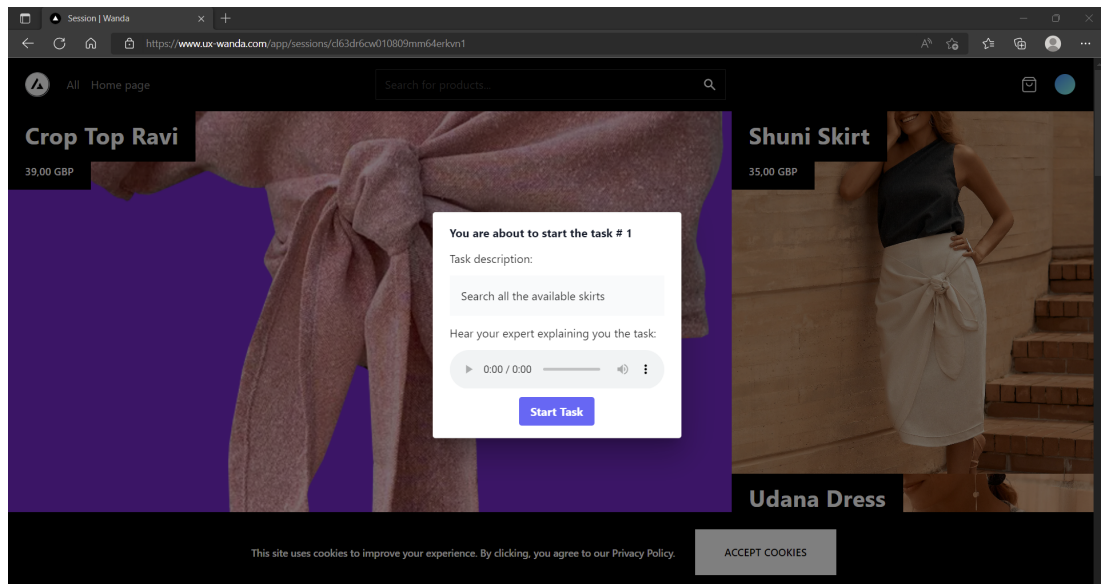


Figure C.9: Participants' study task interface. First iteration of Wanda.

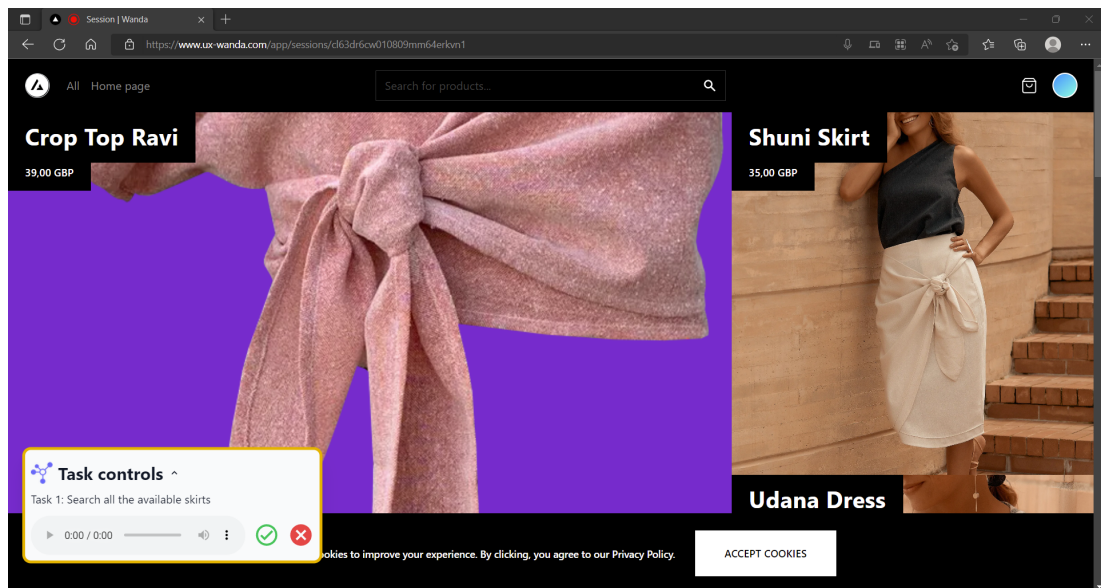


Figure C.10: Participants' task controls. First iteration of Wanda.

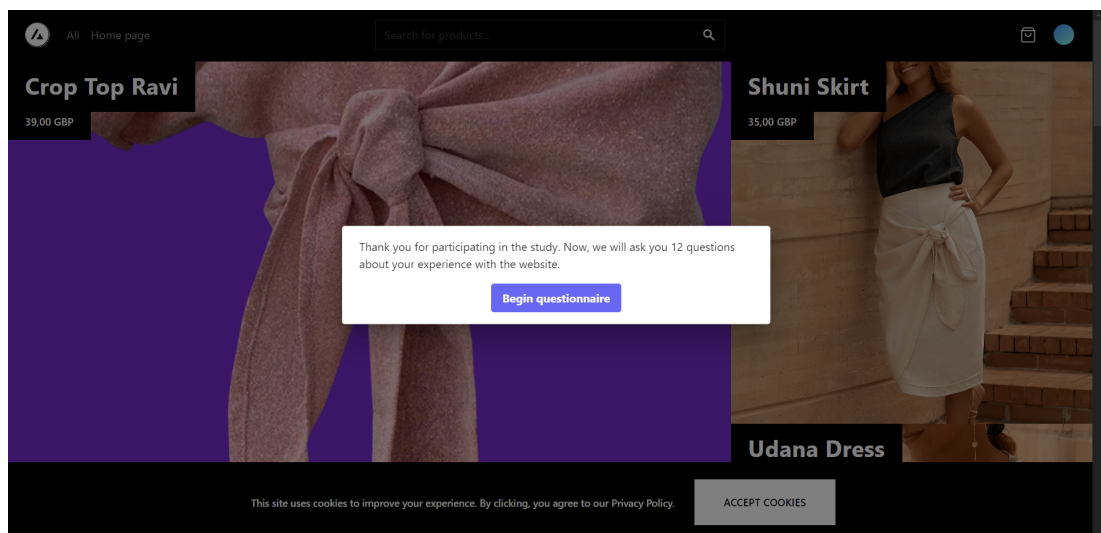


Figure C.11: Participants' begin questionnaire. First iteration of Wanda.

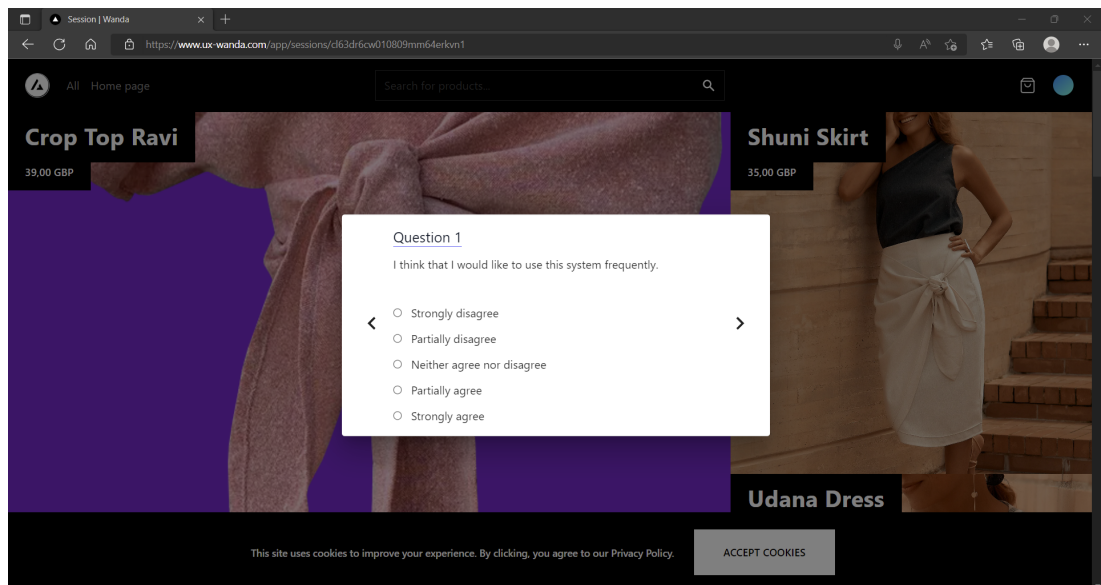


Figure C.12: Participants' question. First iteration of Wanda.

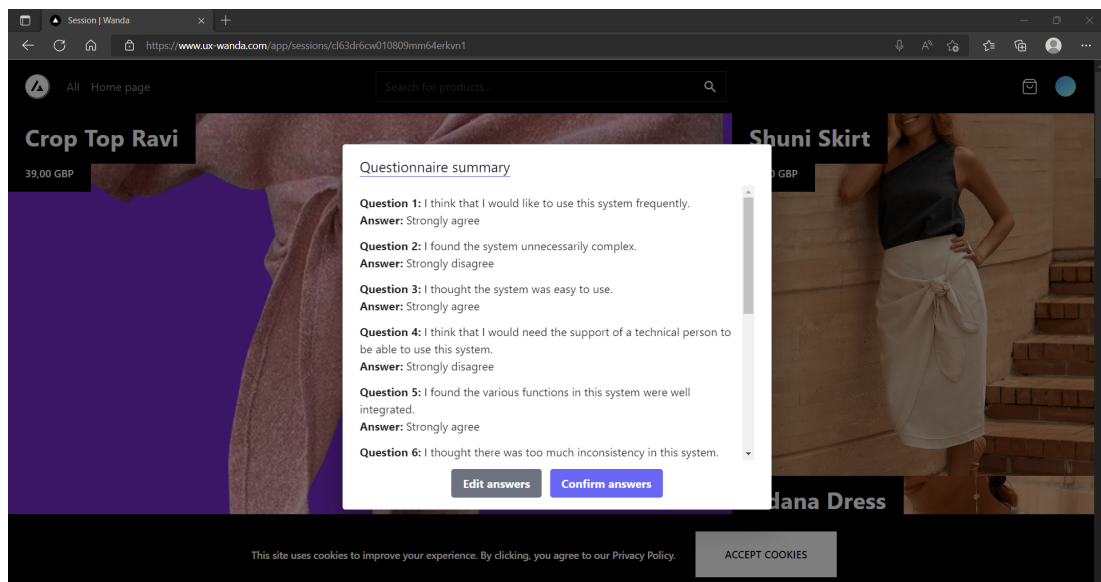


Figure C.13: Participants' questionnaire summary. First iteration of Wanda.

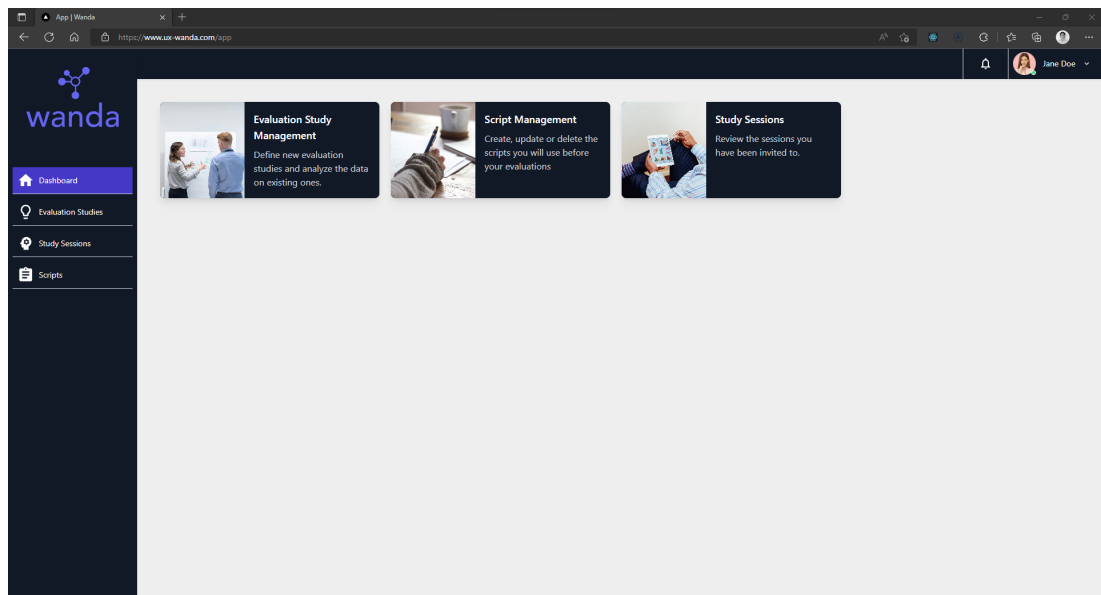


Figure C.14: Experts' home interface. First iteration of Wanda.

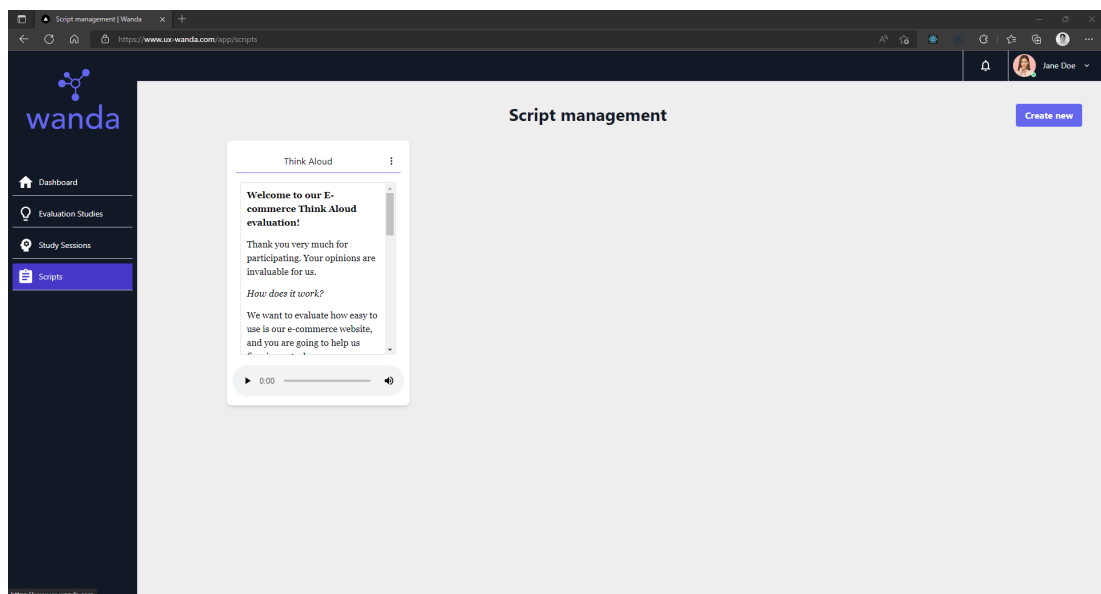


Figure C.15: Experts' scripts interface. First iteration of Wanda.

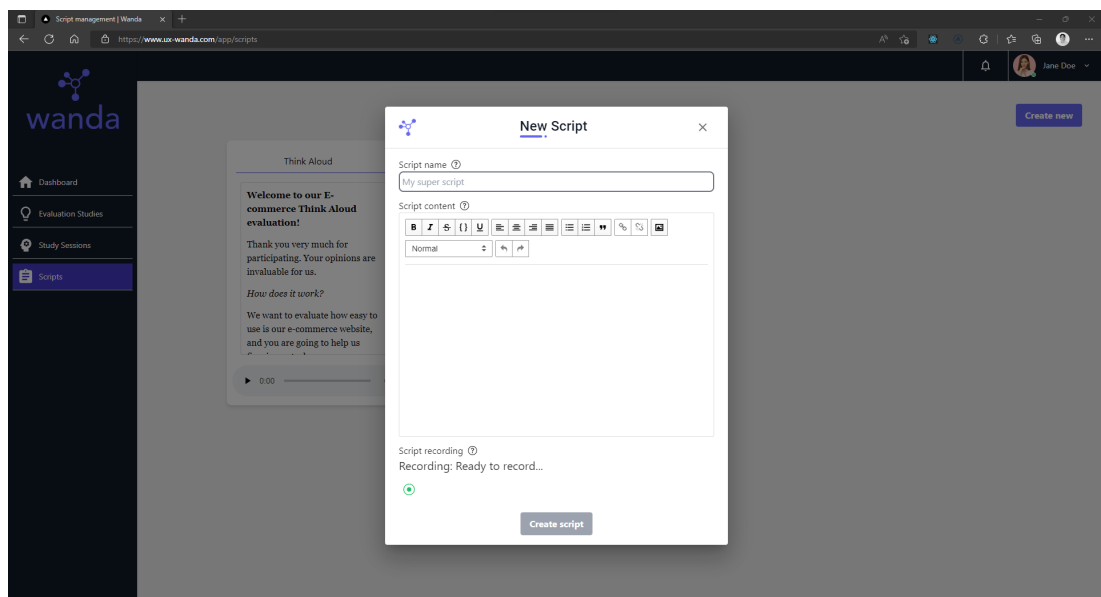


Figure C.16: Experts' new script interface. First iteration of Wanda.

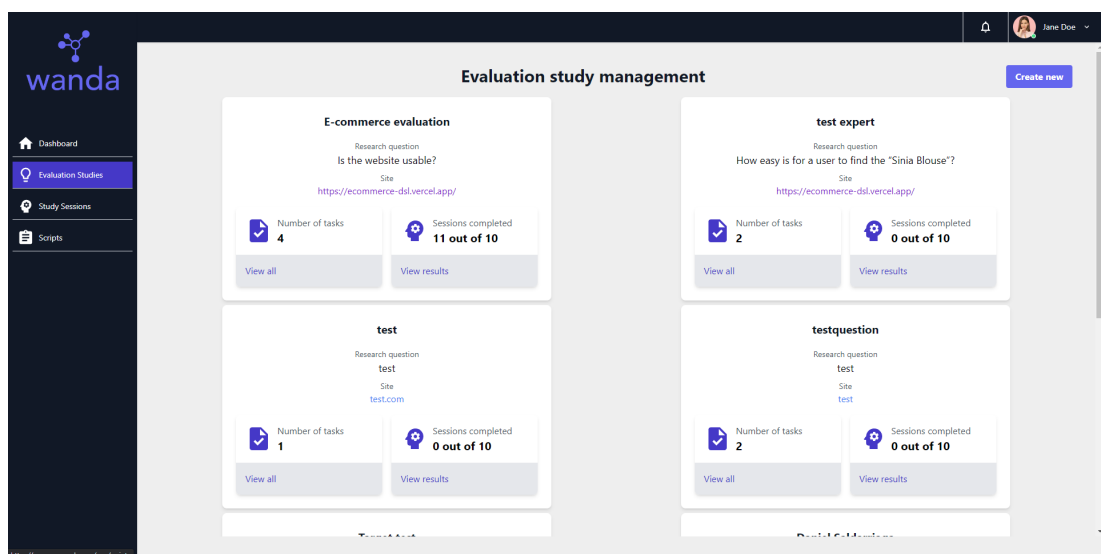


Figure C.17: Experts' evaluation studies interface. First iteration of Wanda.

The screenshot shows a web browser window with the URL <https://www.ux-wanda.com/app/studies/new>. The page title is "New Evaluation Study". On the left is a dark sidebar with the "wanda" logo and navigation links: "Dashboard", "Evaluation Studies", "Study Sessions", and "Scripts". The main content area has a form with the following fields:

- Study name:** MyWebsite
- Target number of participants:** 10
- Research question:** What do I want to find?
- Website:** <https://www.ux-wanda.com>
- Script:** Select a script (dropdown menu)

Below the form is a "Tasks" section with an "Add task" button. A task card titled "Task #1" is visible, containing:

- Task description:** task
- Task url:** <https://www.ux-wanda.com>
- Recording:** Ready to record...

Figure C.18: Experts' new evaluation study form. First iteration of Wanda.

The screenshot shows the "Questionnaire" section of the Wanda application. It features a "System Usability Scale" section with 10 questions and an "Additional questions" section with an "Add" button.

System Usability Scale

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

Additional questions Add

Question 1
question

Figure C.19: Experts' new evaluation study form continuation. First iteration of Wanda.

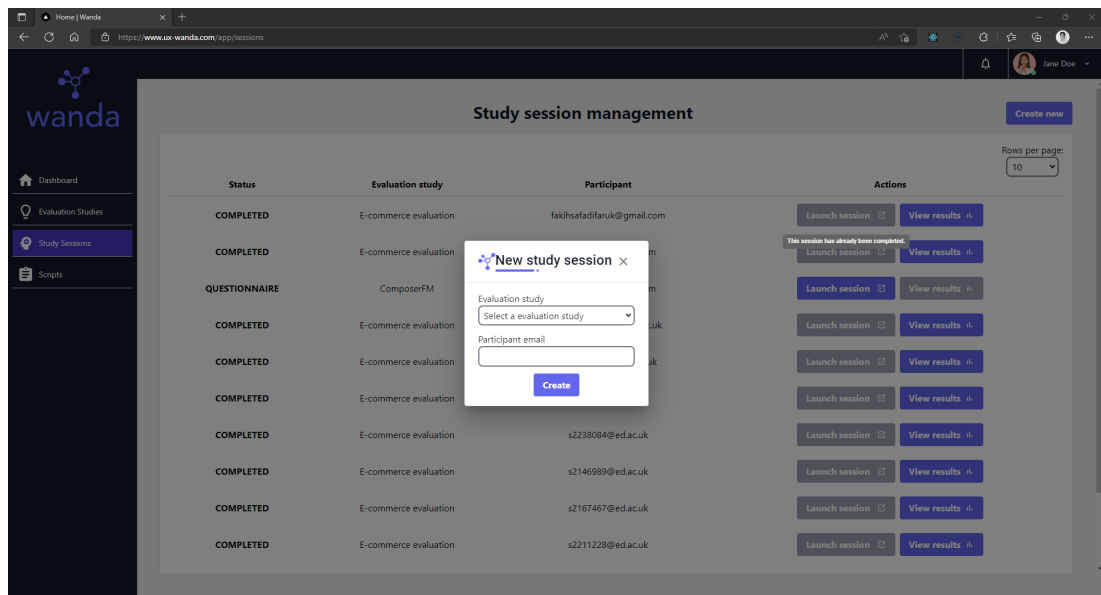


Figure C.20: Experts' new study session form. First iteration of Wanda.

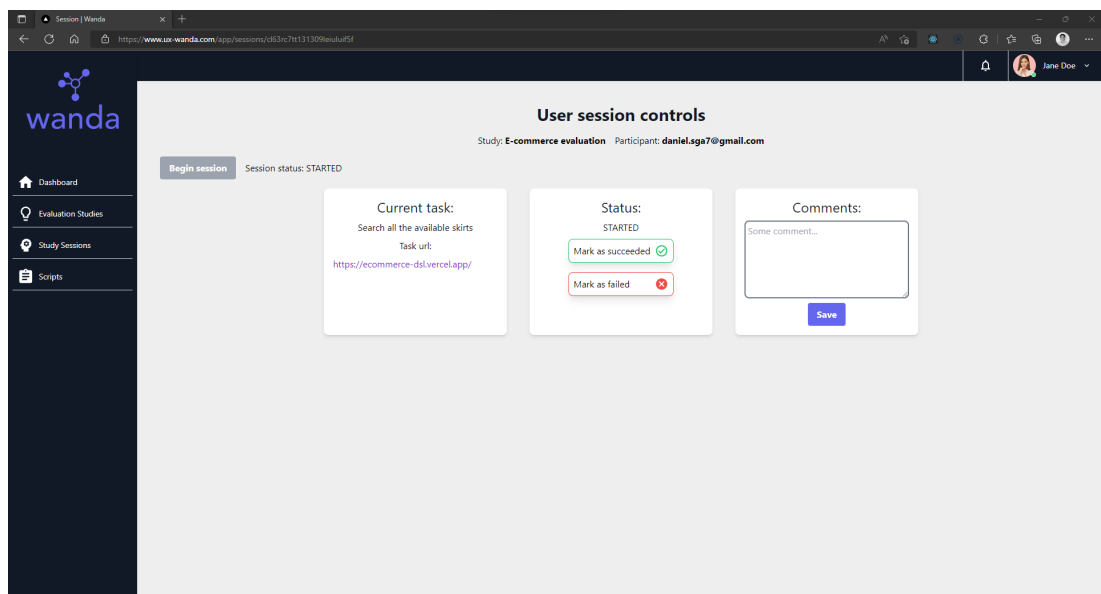


Figure C.21: Experts' study session interface. First iteration of Wanda.

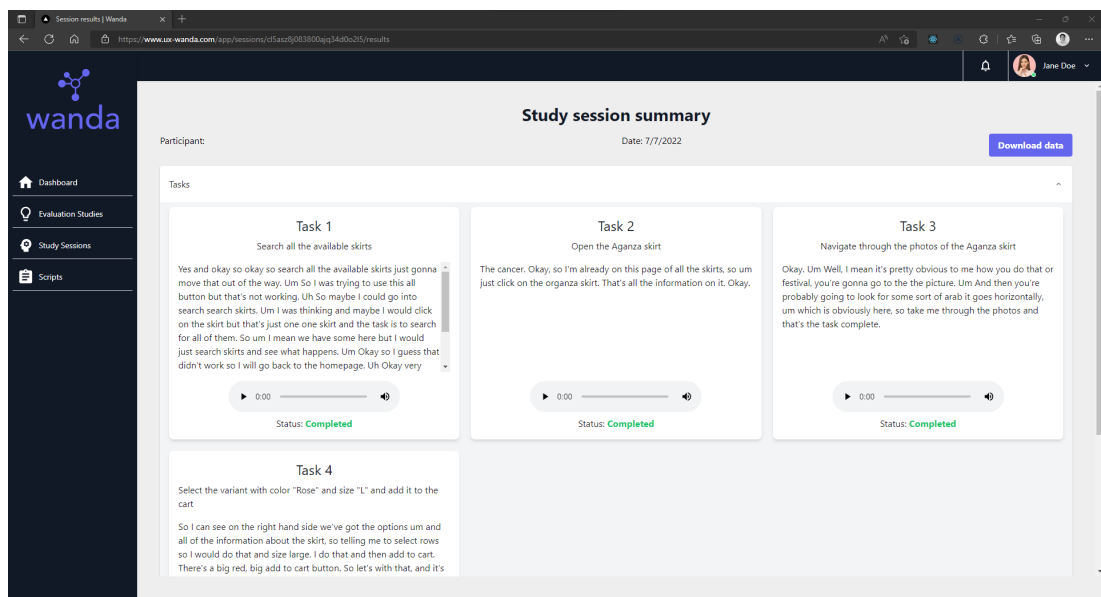


Figure C.22: Experts' study session result interface. First iteration of Wanda.

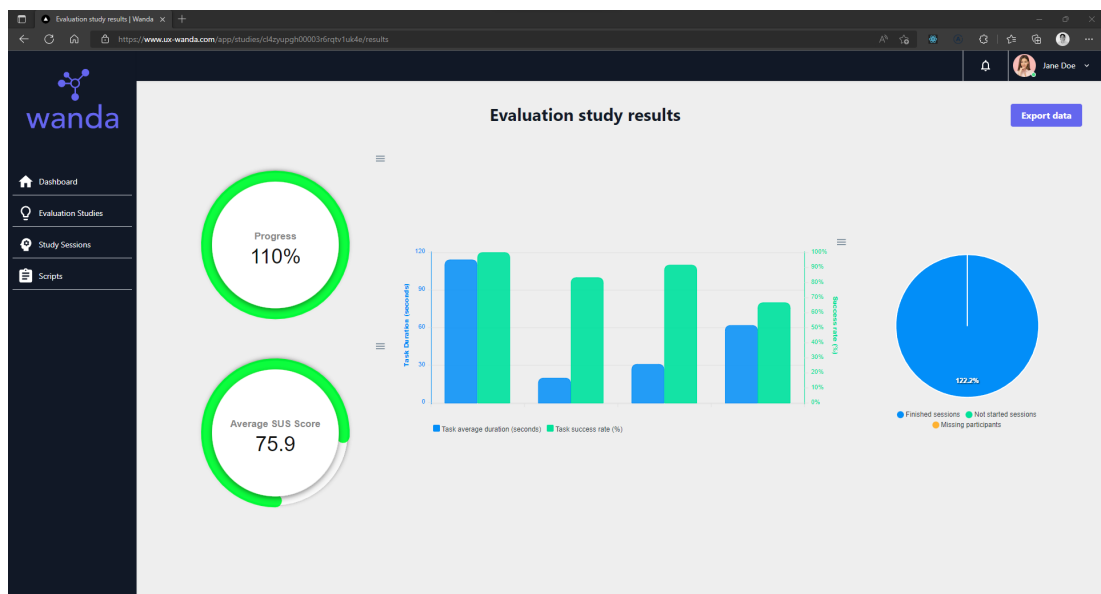


Figure C.23: Experts' evaluation study result interface. First iteration of Wanda.

Appendix D

Formative evaluation

D.1 Participant recruitment

We recruited three experts from The University of Edinburgh, one of which also participated in the requirement gathering study. The two additional experts were also lecturers in the HCI field and had broad academic experience with usability theories.

We also recruited ten students from The University of Edinburgh for acting as participants. Since the participants do not require specific experience with the Think Aloud or Question-Asking Protocol methods, we did not constrain the participants from being from a specific Informatics degree. Instead, we recruited different participants from different degrees to be able to get more feedback.

D.2 Data collection method

Similar to the requirement gathering phase, I conducted one-to-one interviews with both the experts and the participants. Similarly, all the sessions were online, also using Teams [108]. The interviews with the experts were 45-minute long, whereas the interviews with the participants took between 20 and 30 minutes.

Similar to the requirement gathering phase explained before, all the sessions were stored on the University's Microsoft SharePoint [70] server, and the transcripts for each of the sessions were generated to ease the analysis process.

D.3 Materials

For this study, I built a Participant Information Sheet that can be found on the G.2. Since the idea for this evaluation was for the experts and participants to use Wanda and deliver feedback on its features, I built two scripts for guiding the sessions. The scripts are presented in the following sections. Both scripts define a set of tasks that the participants needed to execute in the system.

D.3.1 Scripts

D.3.1.1 Participants

System usability scale

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

Additional questions

1. Have you ever participated in a think aloud evaluation before?
2. What did you like best about Wanda?
3. What did you like least about Wanda?
4. How would you improve the experience of participating in a Think Aloud session using this system?
5. How easy do you think it would be to execute the think aloud evaluation without an expert using this system?
6. Would you recommend any changes to the design of the participant tool?

D.3.1.2 Experts

Wanda Think Aloud Evaluation

Study prerequisites: Google Chrome & maximized window.

Task #1 – Log in

Objective: Log in to the system.

Task description: Go to the website [Home | Wanda \(ux-wanda.com\)](https://ux-wanda.com) and log in using either your Google account or your e-mail address.

Task #2 – Select your role

Objective: Set your name and your role.

Task description: After you log in, write your name and state that your role is “Expert”.

Task #3 – Create a script

Objective: Create a script for the Think Aloud Sessions

Task description:

Go to the “Scripts” menu and create a script using the following data:

Name of the script:

Script Test

Content of the script:

This is a test intended to show how to use the script creation form.

Recording of the script:

Allow your browser’s microphone access and record yourself reading the content of the script.

Task #4 – Create a Evaluation Study

Objective: Create a evaluation study

Task description:

Go to the “Evaluation studies” menu and create a new evaluation study, using the following data:

Name of the study:

E-commerce Evaluation

Target number of participants:

10

Research Question:

How easy is for a user to find the “Sinia Blouse”?

Website:

<https://ecommerce-dsl.vercel.app/>

Script:

Script Test (this is the one you created before).

Task #1:

Description:

Search all the available blouses

Task URL:

<https://ecommerce-dsl.vercel.app/>

Recording: Record yourself reading the description of the task.

Task #2:

Description:

Open the Sinia Blouse

Task URL:

<https://ecommerce-dsl.vercel.app/es/search?q=blouse>

Recording: Record yourself reading the description of the task.

Questionnaire:

Add the system usability scale and add two additional questions:

What did you like best about the e-commerce

What did you like least about the e-commerce

Task #5 – Create a session with a participant

Objective: Create a session with the participant.

Task description: Go to the “Study sessions” menu and create a new study session with the following data:

Evaluation study:

E-commerce Evaluation (this is the one you created before)

Participant e-mail:

daniel.sga7@gmail.com

Task #6 – Launch the session with the participant

Objective: Begin a study session with the participant

Task description:

Click on “Launch session”. You will be redirected to the session admin panel. When you are there, click the button for beginning the session.

Wait for the participant to finish the study.

Task #7 – See the results of the study

Objective: See the results of the study.

Task description:

Go to the “Study sessions” menu and view the results of the session with the participants.

Task #8 – See the results of the evaluation

Objective: See the results of the whole evaluation.

Task description:

Go to the “Evaluation Studies” menu and view the results of the evaluation.

D.4 Procedure

First, all the experts and participants were contacted by e-mail to check their availability and desire to help with the project. Once they agreed to participate in the study, I sent them the Participant Consent Form, the Participant Information Sheet, a time schedule, and the Teams [108] link to the session.

Once the sessions began, I would introduce the objective of the meeting and the objective of the project. Before I began explaining them their tasks, I made sure that they filled the Participant Consent Form and that they agreed to being recorded. After that I would turn on the recording, and begin by asking them about to go to Wanda and log in using their e-mails.

After they were logged in, I sent them the script in PDF format and asked them to begin executing the tasks, making sure to remind them to speak aloud all their thoughts. In the case of the participants, I also made sure to point out when they should provide feedback about Wanda and when about the example e-commerce site, as it can be confusing for them.

After each evaluation, I moved the video recordings from OneDrive to SharePoint, to avoid the risk of data loss due to OneDirve's retention policy of two months at the time of using the software. Once the videos were safely stored, I would extract the transcript of the videos to be able to analyse them faster.

D.5 Data analysis

In the figures D.1 and D.2, I present the code maps that I used for analysing the data of the requirement gathering phase.

Each of the themes identified had different frequencies throughout the evaluation. In the figure D.3 I present the hierarchy of the themes identified on the interviews conducted with the experts. Similarly, in the figure D.4 I show the same information but for the participants. It can be extracted from both figures that the most liked features are the results' visualisations for the experts and that the tool is intuitive for the participants.

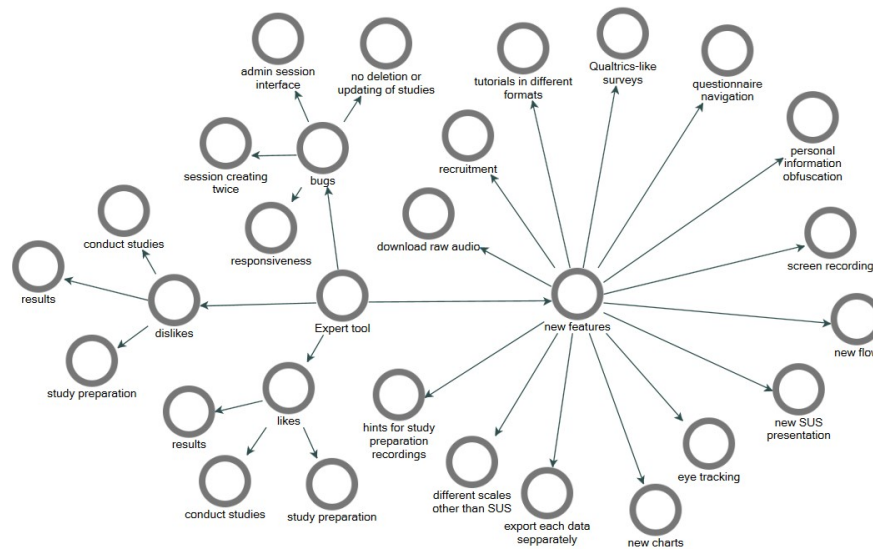


Figure D.1: Formative evaluation gathering themes

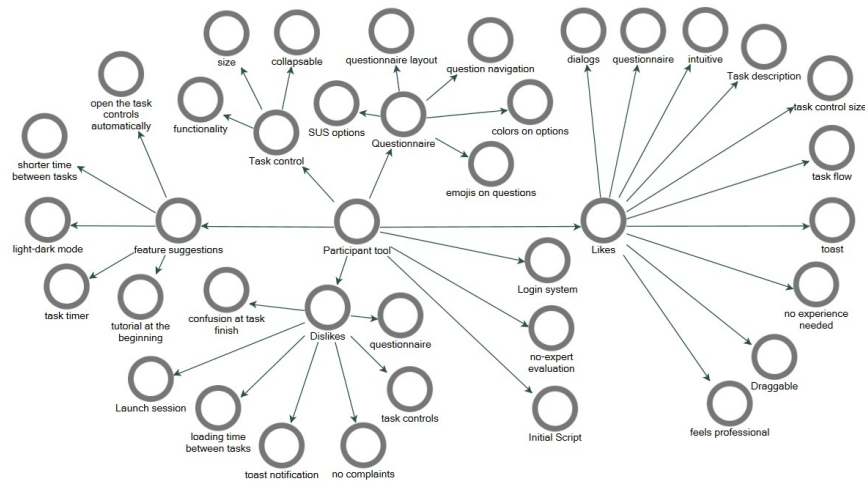


Figure D.2: Formative evaluation gathering themes

D.6 Results

D.6.1 Participant interface

D.6.1.1 Positive points

Overall, the feelings of the participants while using Wanda were quite positive. The most common adjective used to refer to the system was that it feels *professional*, stated by 4 of the 10 participants. Also, six participants commented that Wanda was *intuitive*, using directly that word or a synonym like *easy to use*, *natural* or *easy to learn*.

“I think it was very professionally designed. I mean, it’s a very easy to use tool.” -

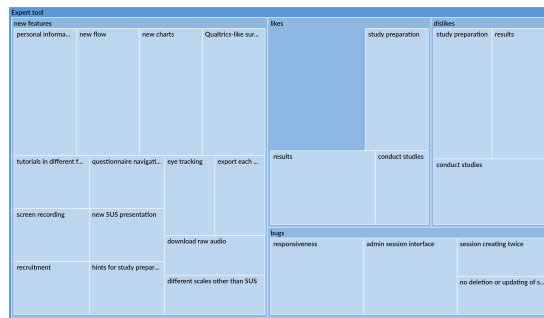


Figure D.3: Experts' formative evaluation theme hierarchy

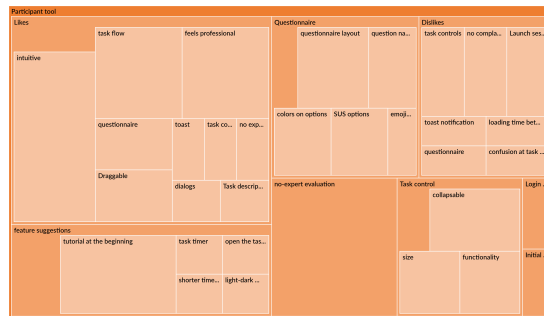


Figure D.4: Participants' formative evaluation theme hierarchy

P10.

When asked about specific features they liked, opinions were diverse. One participant said that they liked the pop-up that alerts that the task status was changed successfully, another said that they felt the questionnaire at the end was *good*, and another liked how the task descriptions were presented. Two participants said they liked how the tool automatically presented the different tasks. Two participants commented on liking how the task controls can be dragged. Two participants said marking a task as succeeded or failed was easy.

D.6.1.2 Points for improvement

I also asked all the participants which features or sections of Wanda they disliked. Again, the opinions were diverse, but two participants complained that the task controls were too small and that they prevented them from seeing important parts of the screen. Another two participants said that they had difficulties finding the *launch session* button, as it was too small.

“When I click into study sessions, maybe the launch session button could be [...]

maybe [a] bigger button” - P4.

Contrary to another participant, one of them said that they did not like the notification pop-up, as it took too long to disappear. There was also one participant confused about the task status, as it was not clear for them when to actually finish the task, because they thought it depended on time more than on the task itself.

Finally, there were two participants who said that they did not have complaints.

D.6.1.3 No-expert evaluation

All the participants were asked about the possibility of doing an evaluation on their own, instead of having an expert present. From the ten participants interviewed, nine said they felt they would be available to do the session on their own and one participant did not comment on the topic. However, their reactions to answering *yes* were mixed. Three of those nine participants doubted when answering *yes*, so they had to think about the answer. Two participants said that they would be able to conduct the session on their own, but only if the instructions were clear. The rest of the participants said that they would definitely be able to use the system on their own.

“Yes, I do think I would be able [to conduct a session without an expert]. Yeah, quite easily” - P4.

One participant additionally commented that they felt that the system would allow experts to gather many participants and increase the audience of the research because the system is easy to use.

D.6.1.4 New feature suggestion

Participants were queried about which new features they would like to see on the system. Four of the ten participants said that the most important feature that Wanda lacks is a tutorial at the beginning, that allows you, as a participant, to get familiar with the mechanics of executing a task.

“Maybe like at the beginning of the evaluation, perhaps there’s like a little tutorial, that highlights the different things that you can do with the system ” - P2.

The rest of the participants commented on different features like a timer for the tasks, a dark mode or shorter loading times between the tasks.

D.6.1.5 System usability scale

At the end of the evaluation, all the participants were asked to fill the System Usability Scale for Wanda. The results are presented in figures D.5 and D.6, showing that participants gave an average score of 87.2 to Wanda, which poses it on the *acceptable* range.

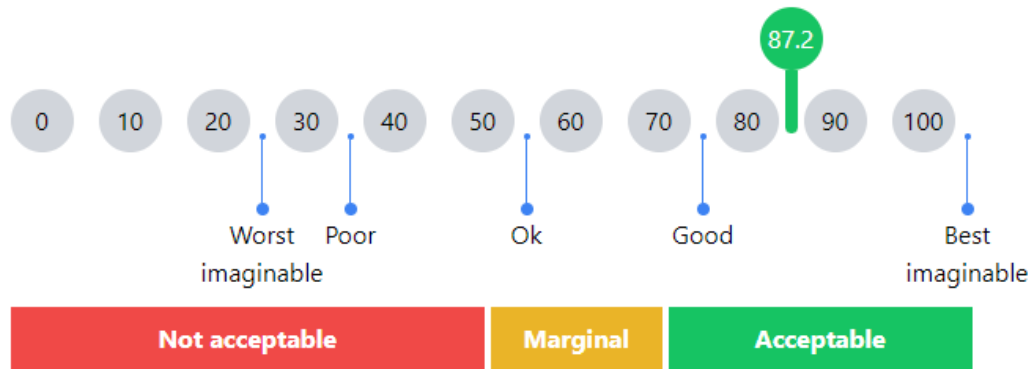


Figure D.5: SUS Results for participants' formative evaluation. Self-drafted image based on [126]

	I found the system unnecessarily complex	I think that this system makes the process of participating in a Think Aloud session easy	I think that I would need support from a technical person every time I use this system	I thought the system was easy to use	I thought there was too much inconsistency in this system	I found the various functions in this system were well integrated	I found the system very cumbersome to use	I would imagine that most people would learn to use this system very quickly	I needed to learn a lot of things before I could get going with this system	I felt very confident using the system	sus
P1	Strongly disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	95
P2	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Disagree	Agree	Strongly disagree	Strongly agree	Neutral	Strongly agree	90
P3	Disagree	Agree	Strongly disagree	Agree	Disagree	Agree	Disagree	Agree	Strongly disagree	Strongly agree	82.5
P4	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Disagree	Strongly agree	97.5
P5	Disagree	Agree	Strongly disagree	Agree	Strongly disagree	Neutral	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	87.5
P6	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	Neutral	Strongly agree	Strongly disagree	Strongly agree	92.5
P7	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Disagree	Disagree	Strongly agree	Strongly disagree	Strongly agree	90
P8	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	97.5
P9	Disagree	Agree	Neutral	Agree	Disagree	Agree	Disagree	Neutral	Disagree	Agree	70
P10	Disagree	Agree	Agree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	70

Figure D.6: SUS Results for participants' formative evaluation.

D.6.2 Expert interface

D.6.2.1 Positive points

As stated before, three experts were interviewed using the Think Aloud [57] method. All three experts could successfully create a new evaluation study, invite a participant, conduct the study and then analyse the results. During that process, all three experts provided excellent feedback on how to improve the system.

Overall, the three experts' most liked feature was that the system automatically transcribes the audio from the participant and splits it between the different tasks. Two

experts commented that such a feature would allow them to process the evaluations faster, and all of them agreed that the feature is crucial in taking more advantage of the Think Aloud method.

All the experts also commented that exporting the data was a good feature, as it allows them to visualize all the information in their own formats if they want.

“I really liked exporting the results, the figures and the different charts. You really don’t need to do too much work. I think in terms of that and the transcription of the the participants’ verbal input, [...] I think it is also very useful” - E3.

Regarding the results, the experts had different opinions. Overall, they liked that the system presented results in charts, allowing them to obtain quick insights into the evaluation they were executing. However, their opinions were diverse regarding each specific presentation of the data, which will be discussed in the following sub-section.

D.6.2.2 Points for improvement

As stated before, there were different opinions about visualising the results of an evaluation. For instance, expert one criticised the way the SUS is presented. They said that the SUS should not be presented as a plain number without context and suggested presenting the SUS using Bangor’s [11] proposal for visualising the results. They also said they did not like pie charts, so they recommended changing the chart with the participant distribution.

Expert two also commented that they would have wanted to check which participants had not completed the evaluation directly in the results page. Such feature was available in the *Study sessions* menu, but not on the *results* menu, which they found confusing.

Experts two and three were aligned on an additional topic regarding personal information. Both expressed concerns that the system is exporting the data in a way that it allows easy mapping of which user executed which evaluation, making it easy to track, for example, the audio of each participant. While such a case may be helpful sometimes, they suggested that it would be better to obfuscate that information and prevent the expert from tracking which participant said what due to privacy concerns.

Additionally, by this iteration, Wanda had not included responsive layouts, and all the experts used the system on a minimised window. Naturally, that caused all the experts to complain about Wanda’s layout. They all experienced responsiveness issues, seeing components on top of other components and elements hiding other elements,

which sometimes caused them not to find features that were required by the evaluation. In those cases, I asked them to maximise their browser windows to find the missing buttons.

“Something I noticed from the previous evaluation, [is that] when the screen is small, the layout is completely wrong. It’s completely damaged.” - E3.

Finally, the first expert was also very critical of Wanda’s data flow. It was unclear to them which menu should be used after each, and they commented that they did not initially understand the difference between creating, conducting and analysing evaluations.

D.6.2.3 New feature suggestion

All experts had different opinions on which features Wanda lacked. Experts two and three commented that it would be good to have a Qualtrics-like [97] survey system, where you could have more freedom in terms of the types of questions you add to the questionnaire.

Expert one suggested changing the flow of the menus on the sidebar of Wanda, including three different options: *Design evaluation studies*, *Conduct evaluation studies* and *Analyse evaluation studies*. These options would also need to match the buttons on the home screen and the tiles on each page.

Additionally, as stated before, expert one suggested changing how the SUS is presented and using Bangor’s [11] proposal to visualise the results.

“Actually, I want to see that Bangor [11] comparative scale and see where this system is with an interpretation” - E1.

Expert two expressed that they would like to be able to download the raw audio, as they do not trust automatic transcription services, especially when the interviewees are non-native English speakers or have strong accents.

Expert three also commented that they would like additional channels like screen recording or eye tracking when the participant executes the evaluation.

Finally, expert two also commented that it would be good for participants to see a tutorial before they begin. As seen in the previous section, this suggestion aligns with what four participants said they would have liked.

D.6.2.4 System Usability Scale

At the end of the evaluation, all the experts were asked to fill the System Usability Scale for Wanda. The results are presented in figures D.7 and D.8, showing that experts gave an average score of 81.7 to Wanda, which poses it on the *acceptable* range.



Figure D.7: SUS Results for experts' formative evaluation. Self-drafted image based on [126]

	I found the system unnecessarily complex	I think that this system makes the process of creating a Think Aloud session easy	I think that I would need support from a technical person every time I use this system	I thought the system was easy to use	I thought there was too much inconsistency in this system	I found the various functions in this system were well integrated	I found the system very cumbersome to use	I would imagine that most people would learn to use this system very quickly	I needed to learn a lot of things before I could get going with this system	I felt very confident using the system	SUS
P1	Strongly disagree	Strongly agree	Strongly disagree	Agree	Disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	92.5
P2	Disagree	Neutral	Disagree	Agree	Neutral	Disagree	Disagree	Agree	Disagree	Neutral	62.5
P3	Strongly disagree	Strongly agree	Disagree	Agree	Strongly disagree	Agree	Strongly disagree	Strongly agree	Disagree	Strongly agree	90

Figure D.8: SUS Results for participants' formative evaluation.

D.7 Updated list of requirements

1. Make an explicit differentiation between Question-Asking Protocol [54] and Think Aloud [57] methods.
2. Integrate the different combinations of evaluations (participant-only or participant-expert) x (Think Aloud or Question-Asking Protocol)
3. Change the visualisation of the SUS to make it similar to Bangor's [11] proposal.
4. Add an onboarding tutorial for participants.
5. Fix the layout of the questionnaire.

6. Fix the responsiveness of the expert interface.
7. Fix the navigation of the expert interface to make it more clear.
8. Add options to update a study and add indications of the study status.
9. Change the *launch* button to make it more clear.
10. Send an e-mail to the participant for connecting directly to the session.

Appendix E

Second iteration of development

E.1 Participants' interface

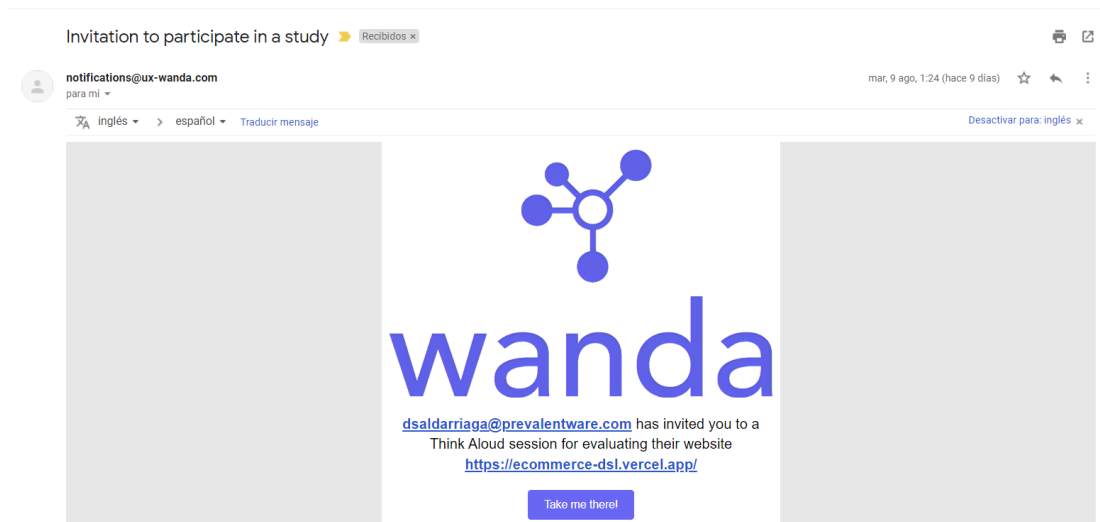


Figure E.1: Email with magic link for connecting to a session. Second iteration of Wanda.

E.2 Experts' interface

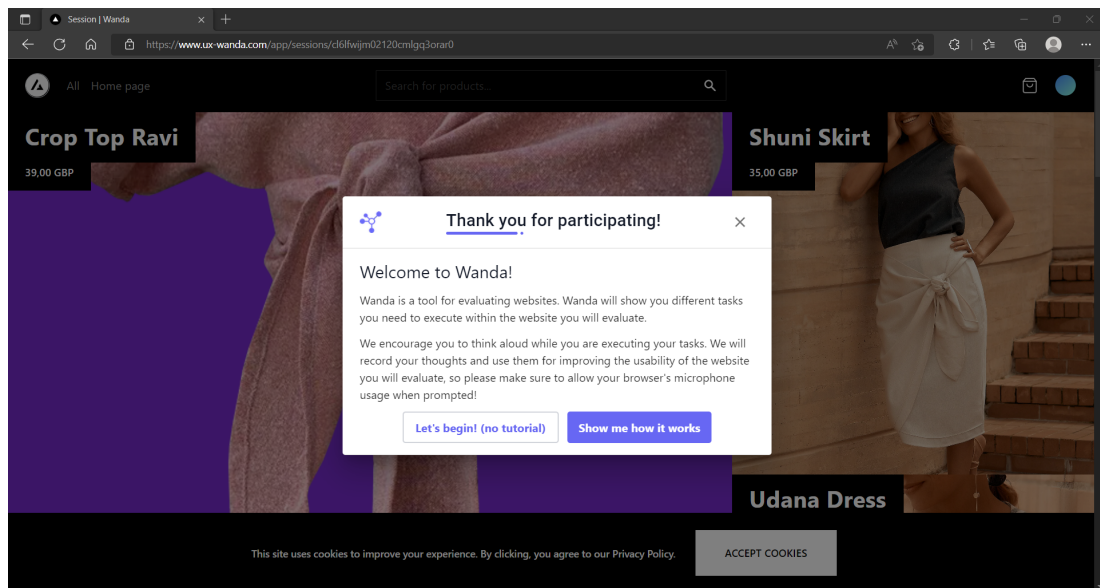


Figure E.2: Initial welcome screen with Think Aloud [57] instructions. Second iteration of Wanda.

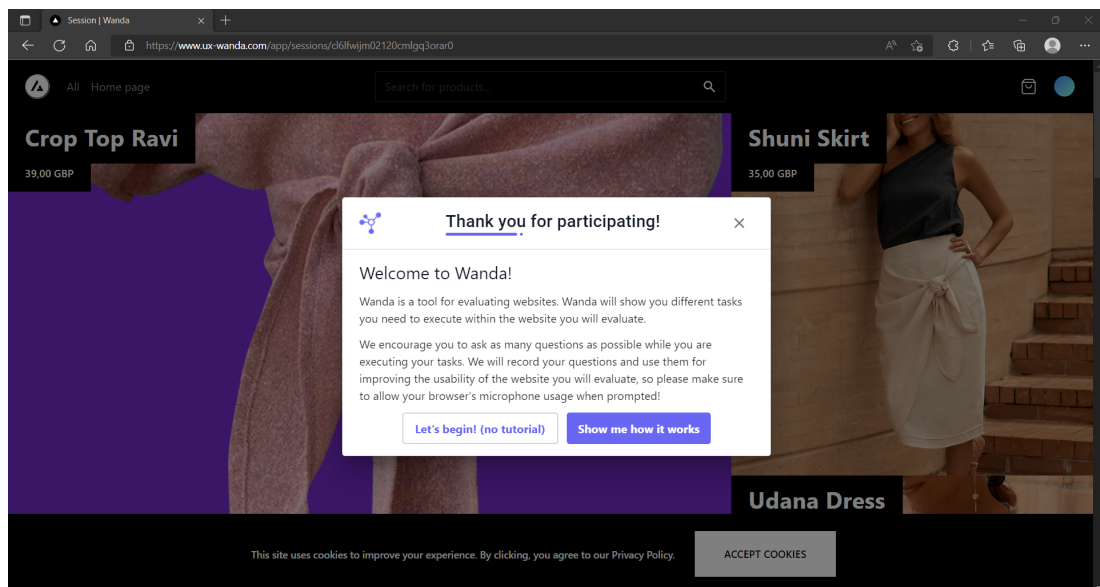


Figure E.3: Initial welcome screen with Question-Asking Protocol [54] instructions. Second iteration of Wanda.

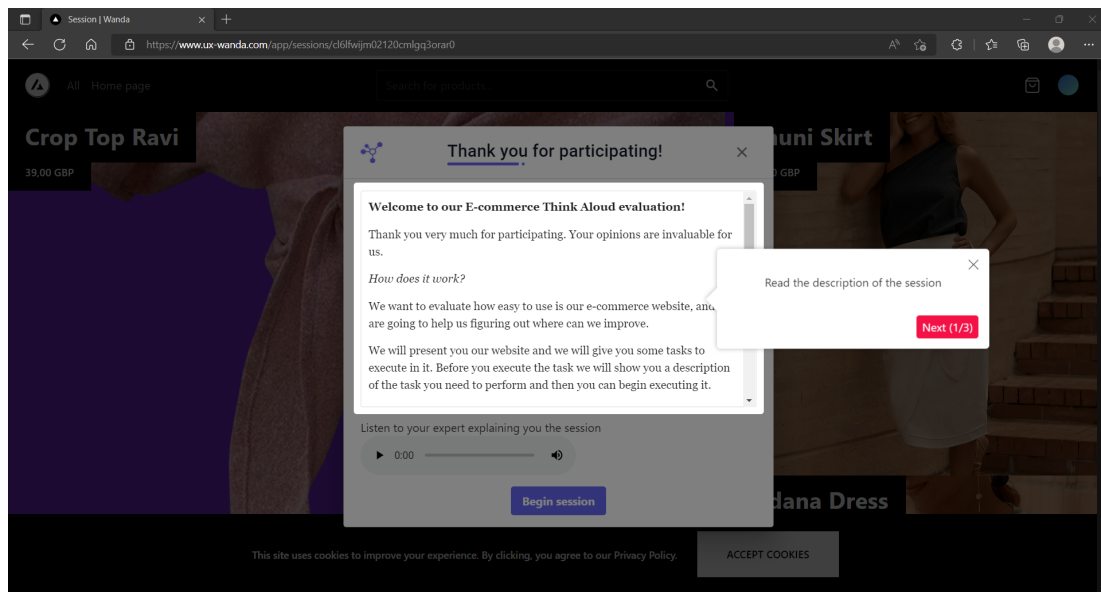


Figure E.4: Tutorial for the participant. Second iteration of Wanda.

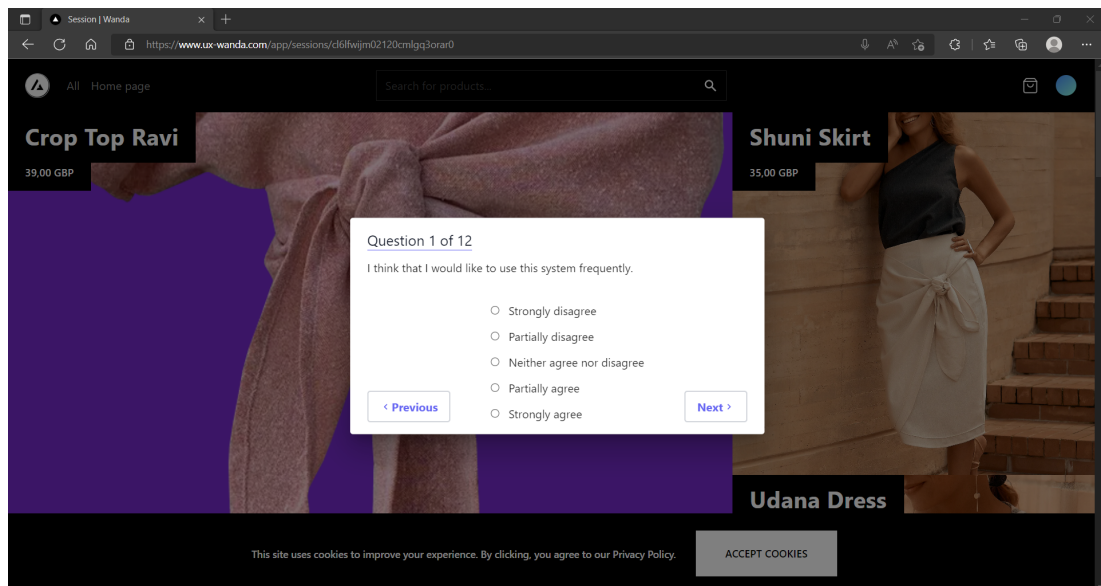


Figure E.5: New questionnaire layout. Second iteration of Wanda.

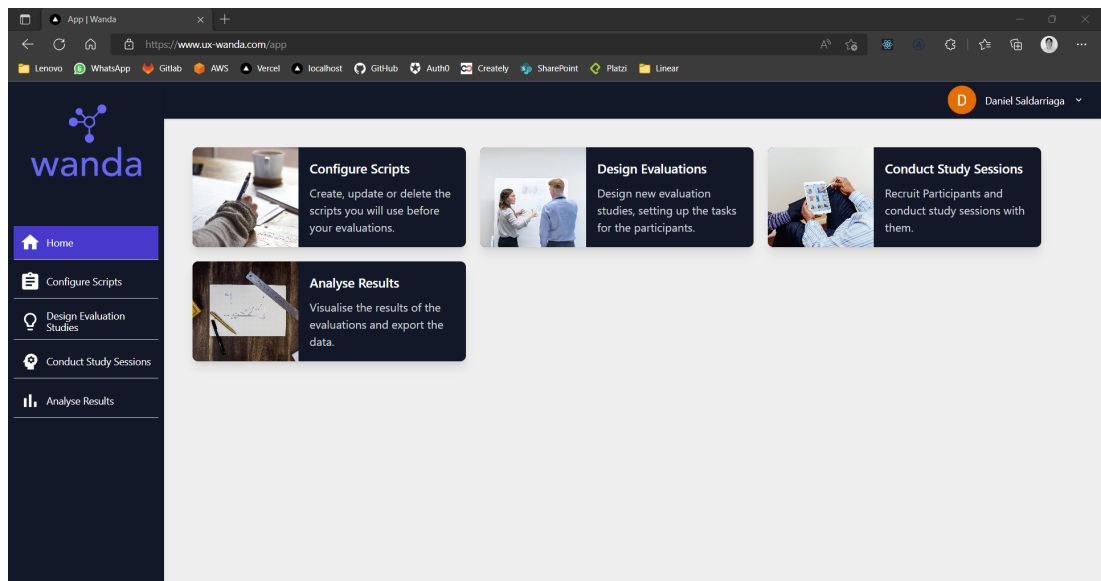


Figure E.6: New home screen for experts. Second iteration of Wanda.

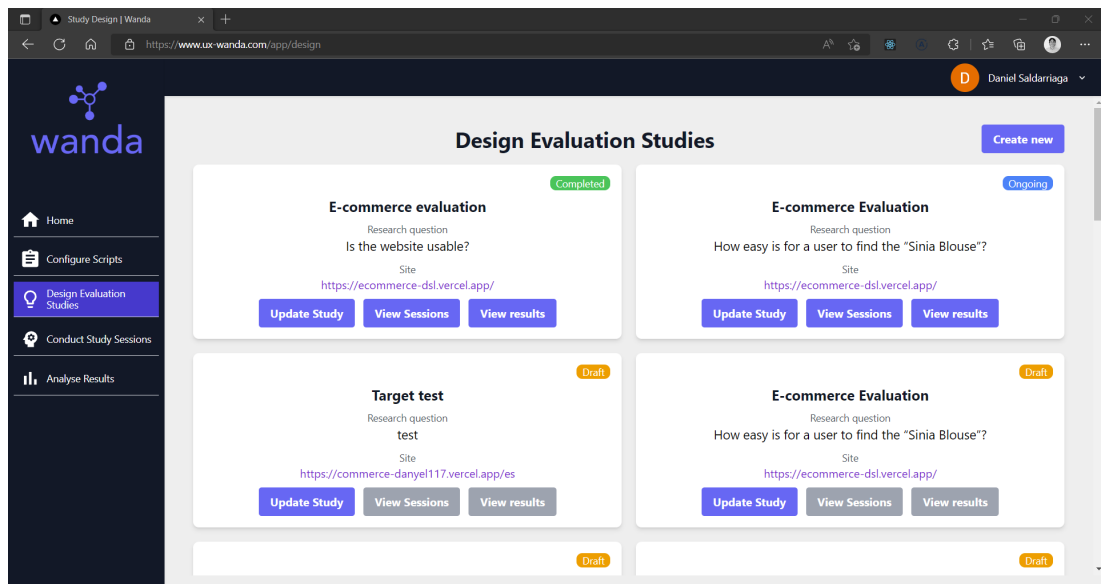


Figure E.7: New menu for evaluation studies. Second iteration of Wanda.

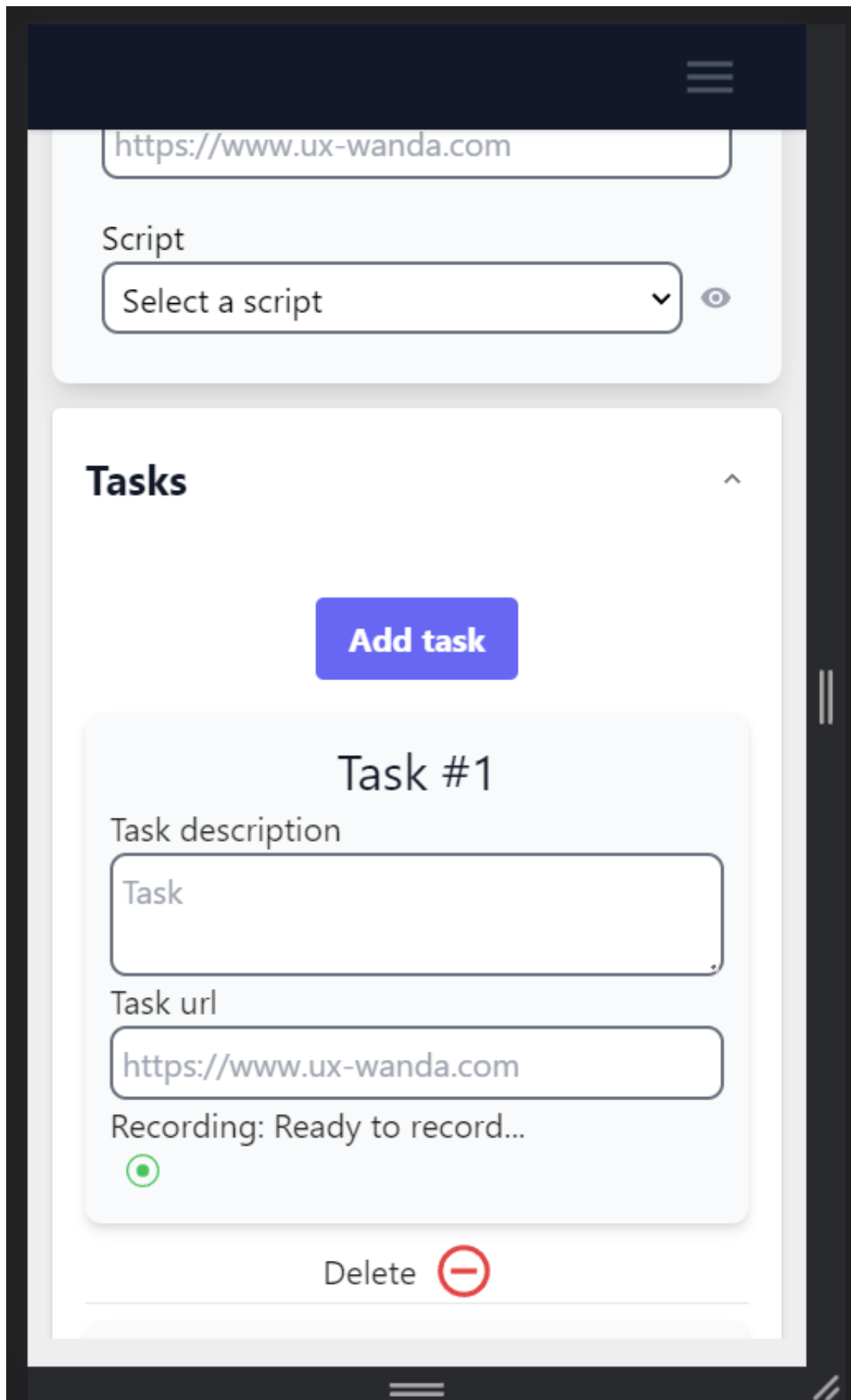


Figure E.8: New responsive screen for creating evaluation studies. Second iteration of Wanda.

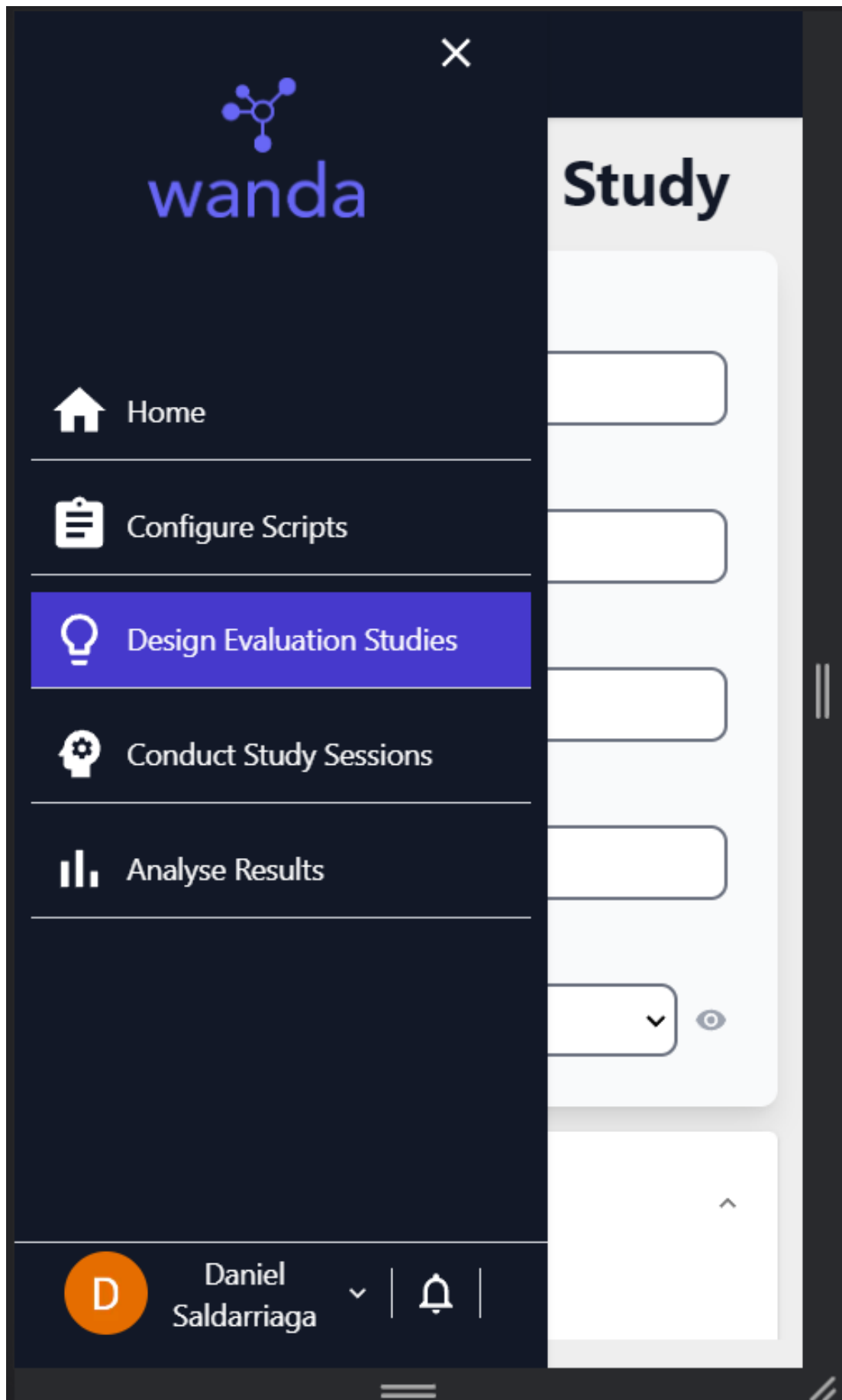


Figure E.9: New responsive sidebar for accessing to the different menus on small screens. Second iteration of Wanda.

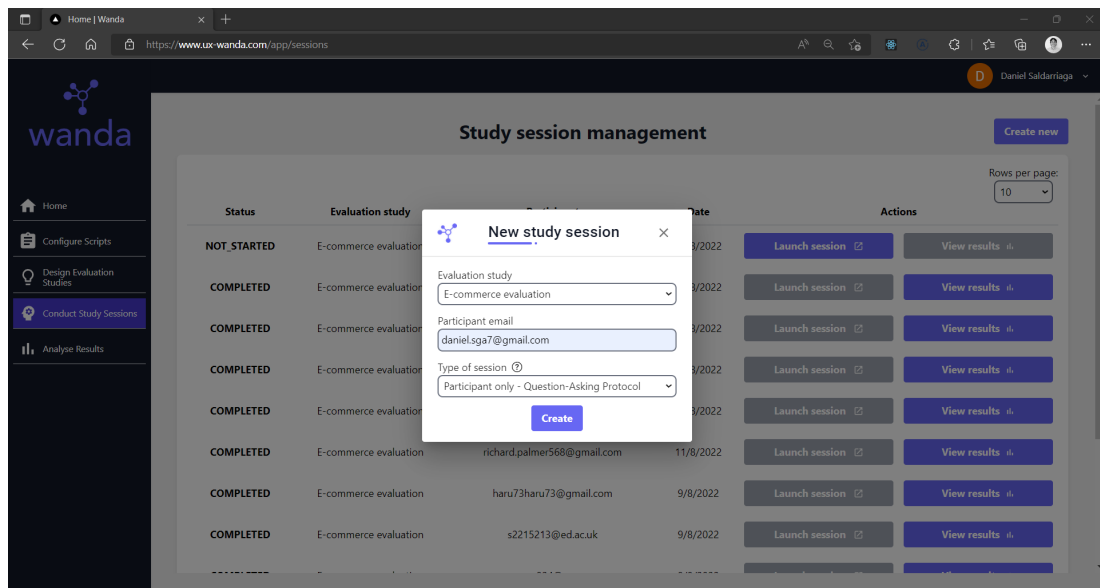


Figure E.10: New menu for creating study sessions including Question-Asking Protocol KatoTakashi1986W. Second iteration of Wanda.

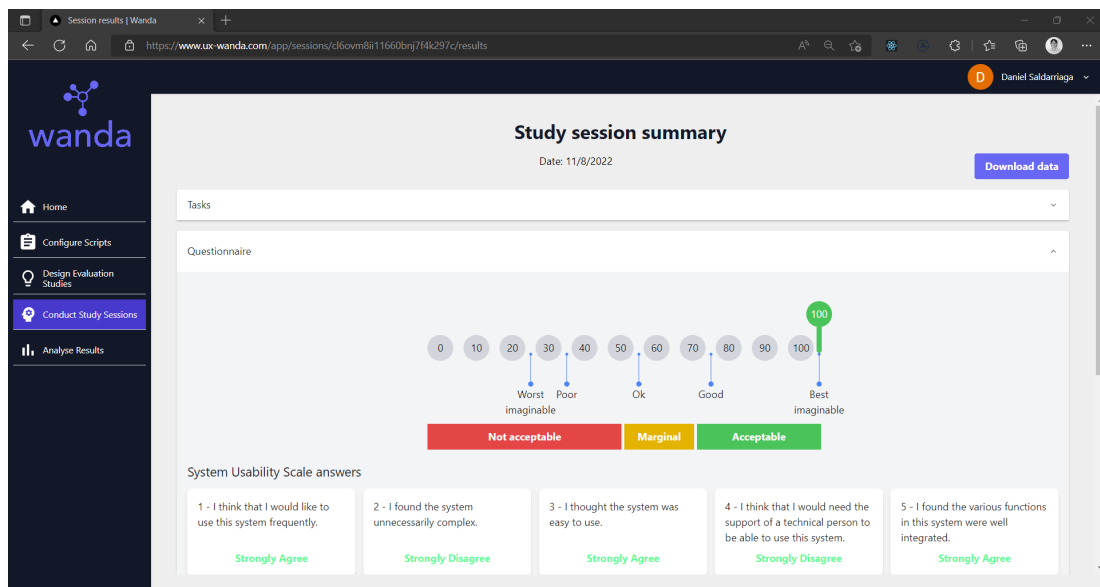


Figure E.11: New menu for analysing study session results with new SUS [16] visualisation. Second iteration of Wanda.

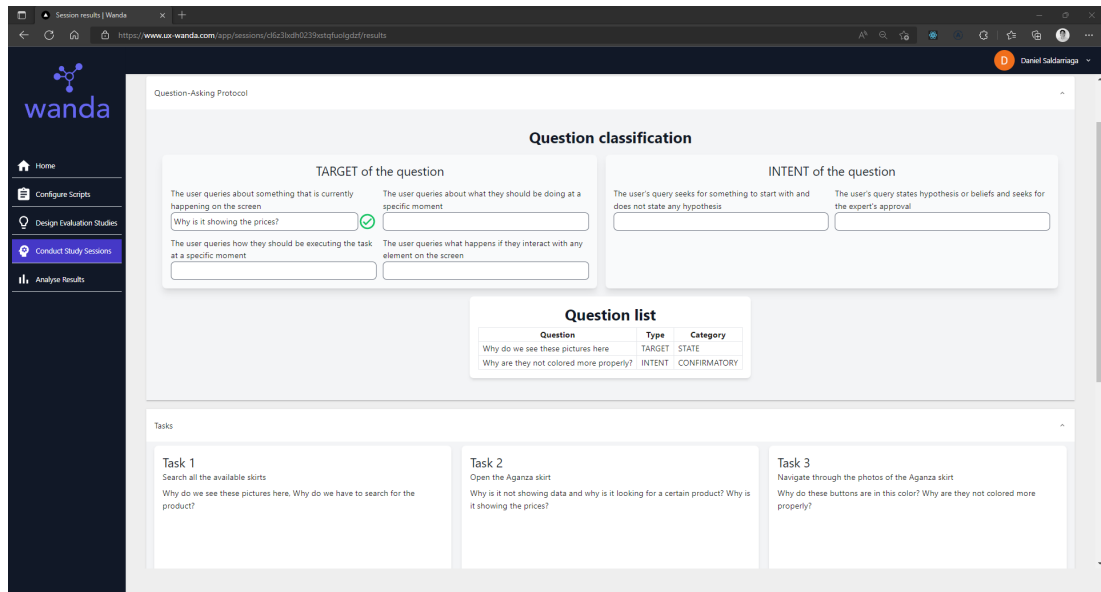


Figure E.12: New for configuring the Question-Asking Protocol [54]. Second iteration of Wanda.

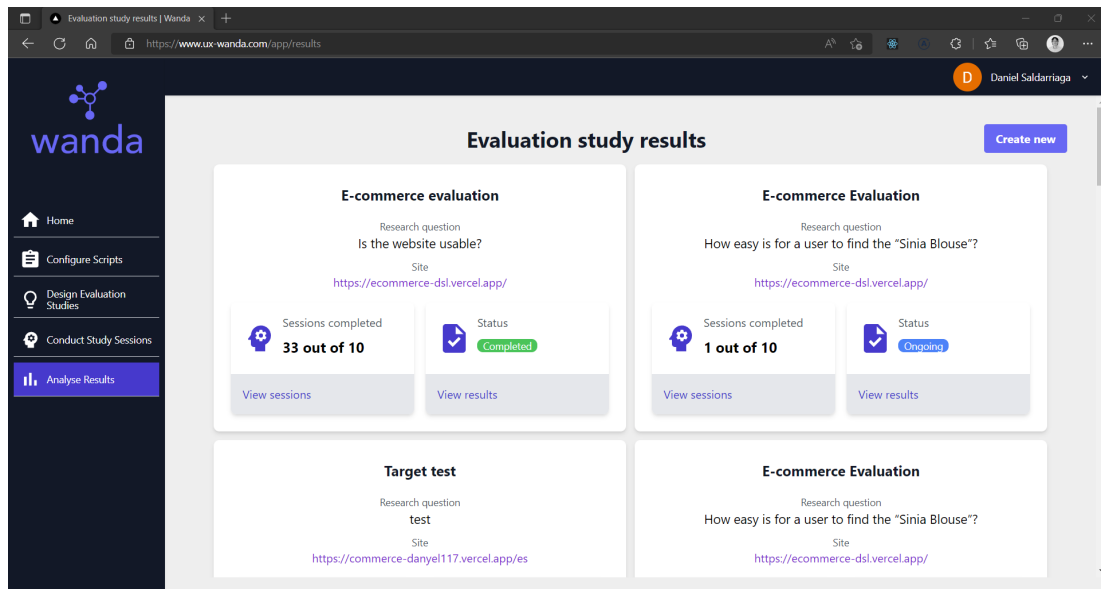


Figure E.13: New menu containing all the results for the evaluation studies. Second iteration of Wanda.

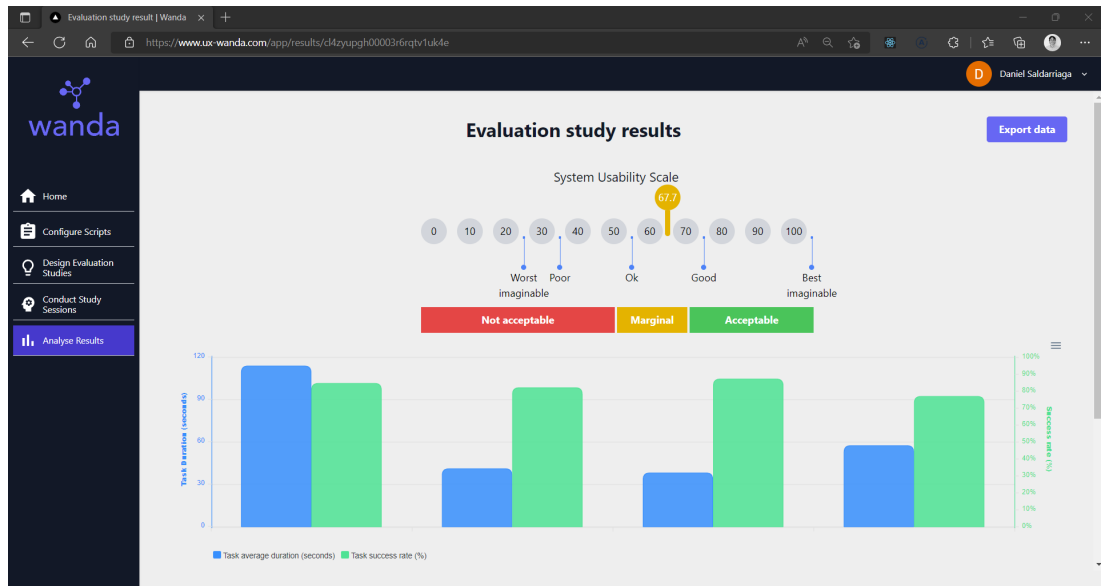


Figure E.14: New dashboard with the result for an evaluation study. Second iteration of Wanda.



Figure E.15: Change in the System Usability Scale visualisation [16]. Second iteration of Wanda.

Appendix F

Summative evaluation

F.1 Participant recruitment

Similar to the formative, this evaluation is splitted into participants and experts. For this study we recruited five HCI experts from The University of Edinburgh. From those five experts, one was present in both the requirement gathering and the formative evaluations, one was only present in the requirement gathering, one was present only in the formative and two were completely new to the project.

On the participant side, we followed a similar method to what we did on the formative evaluation, contacting potential participants through e-mail and social networks. We also difussed two e-mails to all the MSc and undergraduate students of the Informatics department from The University of Edinburgh to try and broaden our participant base.

F.2 Data collection method

Similar to the requirement gathering phase, I conducted one-to-one interviews with the experts. We were able to get 14 potential participants, so doing one-to-one sessions with them was not feasible. For that reason, we conducted three focus groups with the potential participants.

All the sessions were online using Microsoft Teams [108]. The sessions with the experts lasted between 30 and 45 minutes and the focus groups were 45-minute long.

Similar to the evaluations explained before, all the sessions were stored on the University's Microsoft SharePoint [70] server, and the transcripts for each of the sessions were generated to ease the analysis process.

F.3 Materials

For this study, I built a Participant Information Sheet that can be found on the G.3. Similar to the formative evaluation, I built two scripts for guiding the sessions. The scripts can be found in the following sections. Both scripts define a set of tasks that the participants needed to execute in the system.

F.3.1 Scripts

F.3.1.1 Participants

Focus Group – Evaluation of Wanda

Objective: Evaluate the usability of [Wanda \(ux-wanda.com\)](https://www.ux-wanda.com/)

What is Wanda?

Wanda is a tool that allows experts to conduct remote usability studies with their participants. It is a tool where you, as a participant, can use the Think Aloud and Question-Asking Protocol methods to express what you think about a website.

When using Wanda, you will be presented with a set of tasks you need to execute. The idea is that you should speak aloud every thought you have about the application, and the system will record the audio while you are speaking. That audio information is then stored in the system and used by an expert to improve a website's usability.

Today you will take the role of one of those participants and use Wanda to evaluate a generic e-commerce application.

Throughout the evaluation, keep in mind that this focus group's main objective is to gather feedback about **Wanda**, not about the e-commerce. You will just use the e-commerce as an example, but the main goal is understanding how **Wanda** works.

Instructions

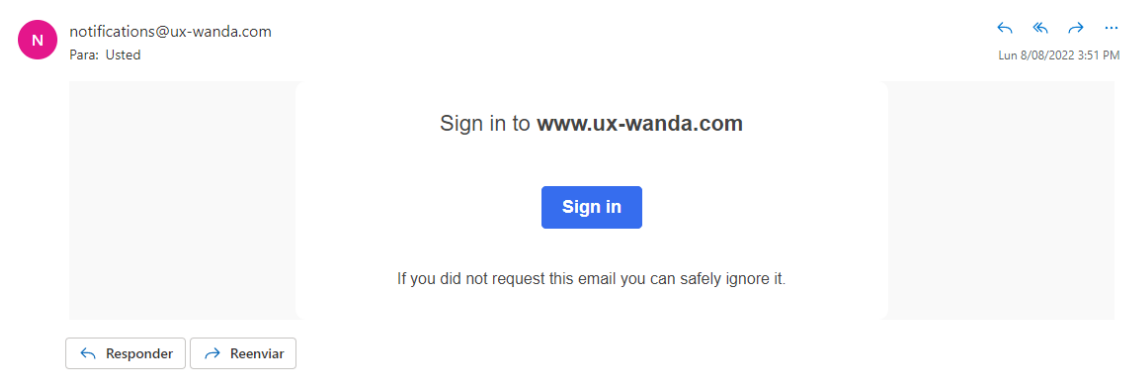
PART 1: Log in to the system

Go to <https://www.ux-wanda.com/>, click "Take me there", and choose your preferred login method.

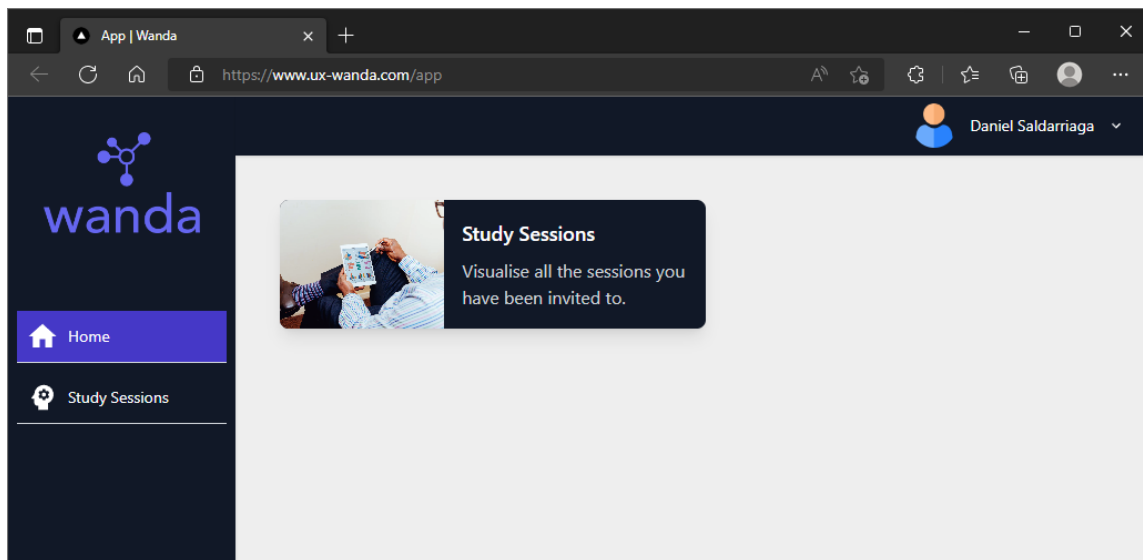
If you choose to log in with Google, follow the instructions on the screen for selecting your Google account.

If you choose to log in with your e-mail, you need to wait for the system to send you an e-mail with a magic link. It may take up to 30 seconds, and it may go to your spam, so please make sure to check it.

Once you receive the e-mail, you can click the "sign in" button and be redirected to Wanda.



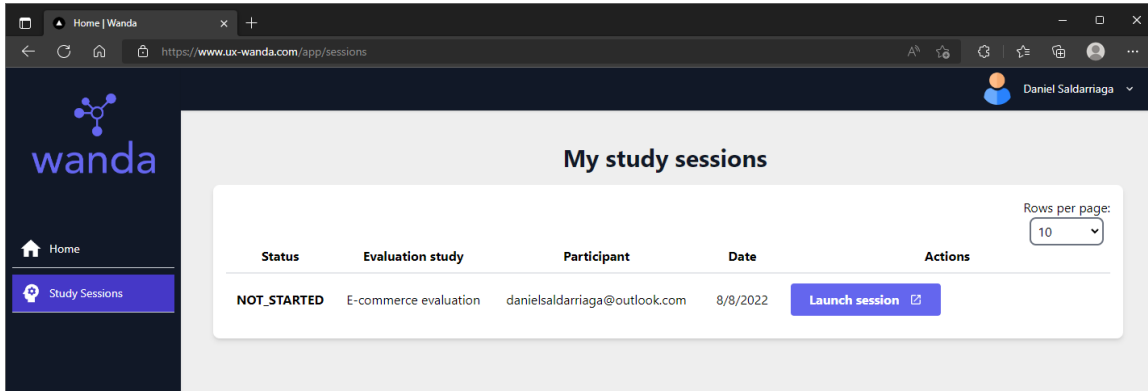
Once you are logged into Wanda, you will see the following screen:



Go to the "Study Sessions" menu and wait for further instructions from the moderator.

PART 2: Start the evaluation

You should see a table with a session you've been invited to. Click "Launch session".

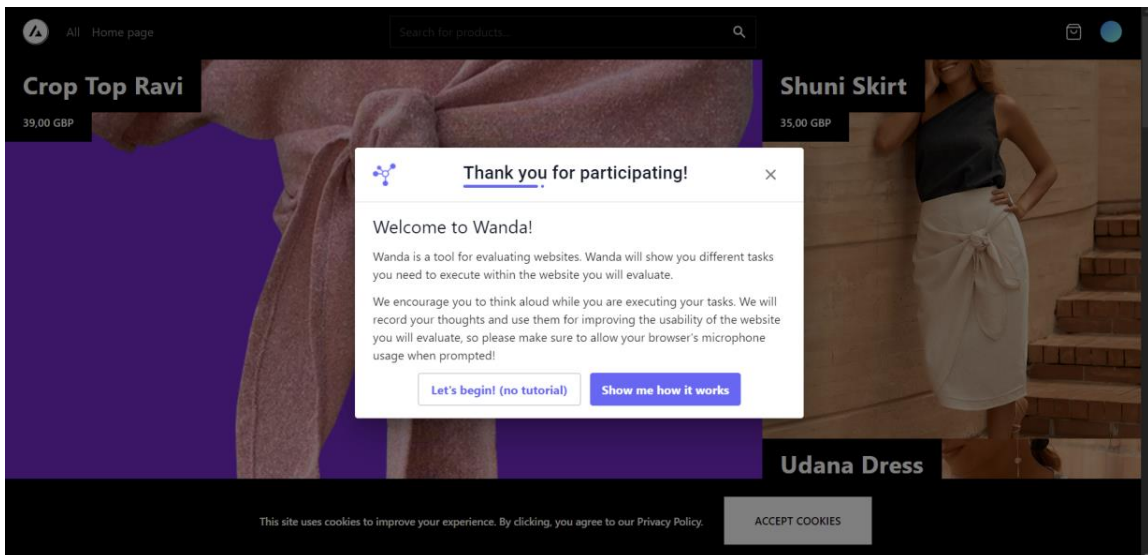


The screenshot shows a web browser window with the URL <https://www.ux-wanda.com/app/sessions>. The page title is "My study sessions". On the left, there is a sidebar with the Wanda logo and navigation links for "Home" and "Study Sessions". The main content area displays a table with the following data:

Status	Evaluation study	Participant	Date	Actions
NOT_STARTED	E-commerce evaluation	danielsaldarriaga@outlook.com	8/8/2022	Launch session

There is a "Rows per page:" dropdown menu set to "10" in the top right corner of the table area.

You will be redirected to the e-commerce and should see the following screen:



The screenshot shows a mock e-commerce website. The background features a purple top with the text "Crop Top Ravi" and "39,00 GBP", a white skirt with the text "Shuni Skirt" and "35,00 GBP", and a white dress with the text "Udana Dress". A central white overlay box contains the following text:

Thank you for participating!

Welcome to Wanda!

Wanda is a tool for evaluating websites. Wanda will show you different tasks you need to execute within the website you will evaluate.

We encourage you to think aloud while you are executing your tasks. We will record your thoughts and use them for improving the usability of the website you will evaluate, so please make sure to allow your browser's microphone usage when prompted!

[Let's begin! \(no tutorial\)](#) [Show me how it works](#)

At the bottom of the page, there is a cookie consent banner: "This site uses cookies to improve your experience. By clicking, you agree to our Privacy Policy." with an "ACCEPT COOKIES" button.

In the background, you see the mock e-commerce we will use as a test. In the foreground, you can see Wanda's instructions.

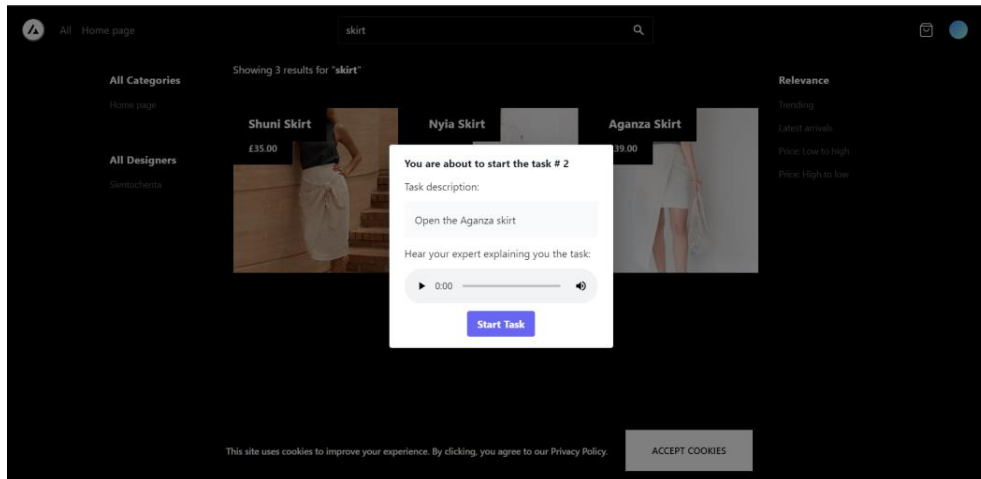
Read the message on the screen and wait for the organiser of the focus group to give further instructions.

PART 3: Follow the tutorial

Click "Show me how it works" and follow all the instructions until you complete the tutorial.

Once you complete the tutorial, you can finish the first task within the e-commerce, which is to look for all the available skirts.

Once you complete the first task, the system will show the second task on the screen, and you will see this window:



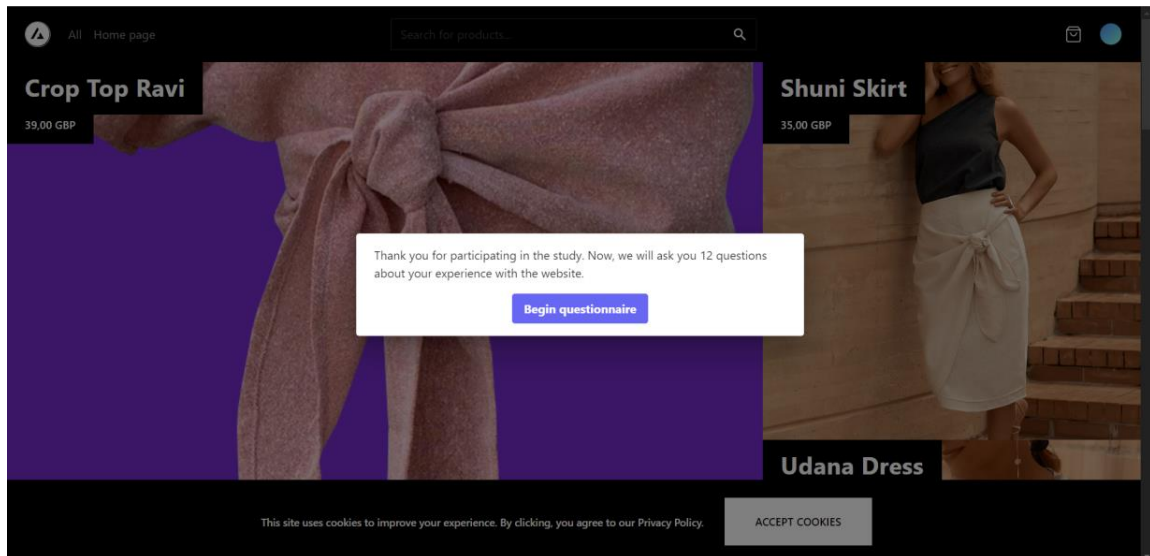
Wait for the organiser of the focus group to give you further instructions.

PART 4: Finish the three tasks on the e-commerce

Begin executing tasks #2, #3 and #4.

The instructions for each task can be seen before executing each task. Also, you can find it in the task controls in the bottom-left corner.

Once you finish the fourth task, the system should show this window:



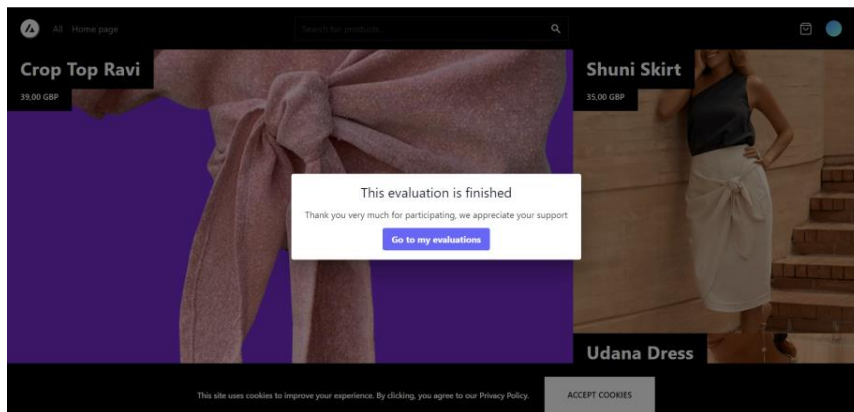
Wait for further instructions from the organiser of the focus group

PART 5: Fill out the questionnaire

Click "begin questionnaire" and answer all the questions. At this stage, your answers to the questionnaire should be about the e-commerce.

*Remember that you are acting as the evaluator of the e-commerce, which means that you are using Wanda for evaluating the **e-commerce**.*

Once you finish the questionnaire, you see the following window:



Click "Go to my evaluations" and wait for instructions from the organiser of the focus group.

PART 6: Freely explore the system

You now have one minute to explore the rest of the platform freely. You can click the different menus, sign out, sign in again and roam through the page.

Once the time finishes, wait for the organiser of the focus group to give you further instructions.

F.3.1.2 Experts

Wanda Think Aloud Evaluation

Task #1 – Log in

Objective: Log in to the system.

Task description: Go to the website [Home | Wanda \(ux-wanda.com\)](https://ux-wanda.com) and log in using either your Google account or your e-mail address.

Task #2 (Only required if the user is not registered) – Select your role

Objective: Set your name and your role.

Task description: After you log in, write your name and state that your role is “Expert” if the system

Task #3 – Create a script

Objective: Create a script for the Think Aloud Sessions

Task description:

Go to the “Configure Scripts” menu and create a script using the following data:

Name of the script:

Think Aloud Explanation

Content of the script:

Welcome to our E-commerce Think Aloud evaluation!

Thank you very much for participating. Your opinions are invaluable for us.

How does it work?

We want to evaluate how easy to use is our e-commerce website, and you are going to help us figuring out where can we improve.

We will present you our website and we will give you some tasks to execute in it. Before you execute the task, we will show you a description of the task you need to perform and then you can begin executing it.

The idea is simple: speak aloud all your thoughts! Everything that comes to your mind is helpful for us and we will improve our website based on your comments.

Do not worry if you find a task very difficult, or even if you cannot complete it. That is precisely the type of feedback we require.

During the execution of each task, you will have some controls for playing the description of the task or for marking a task as succeeded or failed.

Enjoy your evaluation and thank you very much in advance!

Recording of the script:

Allow your browser's microphone access and record yourself reading the content of the script.

Task #4 – Create a Evaluation Study

Objective: Create a evaluation study

Task description:

Go to the “Design Evaluation studies” menu and create a new evaluation study, using the following data:

Name of the study:

E-commerce Evaluation – Summative test

Target number of participants:

10

Research Question:

How easy is for a user to find the “Sinia Blouse”?

Website:

<https://ecommerce-dsl.vercel.app/>

Script:

Think Aloud Explanation (this is the one you created before).

Task #1:

Description:

Search all the available blouses

Task URL:

<https://ecommerce-dsl.vercel.app/>

Recording: Record yourself reading the description of the task.

Task #2:

Description:

Open the Sinia Blouse

Task URL:

<https://ecommerce-dsl.vercel.app/es/search?q=blouse>

Recording: Record yourself reading the description of the task.

Questionnaire:

Add the system usability scale and add two additional questions:

What did you like best about the e-commerce?

What did you like least about the e-commerce?

Task #5 – Create a session with a participant

Objective: Create a session with the participant.

Task description: Go to the “Conduct Study sessions” menu and create a new study session with the following data:

Evaluation study:

E-commerce Evaluation – Summative test (this is the one you created before)

Participant e-mail:

daniel.sga7@gmail.com

Evaluation type:

Participant only – think aloud

Task #6 – See the session of the participant

Objective: Check the participant interface

Task description:

Wait for the participant to finish the study.

Task #7 – See the results of the study

Objective: See the results of the study.

Task description:

Go to the “Study sessions” menu and view the results of the session with the participants.

Task #8 – See the results of the evaluation

Objective: See the results of the whole evaluation.

Task description:

Go to the “Analyse Results” menu and view the results of the evaluation.

F.4 Procedure

First, all the experts and participants were contacted by e-mail to check their availability and desire to help with the project. Once they agreed to participate in the study, I sent them the Participant Consent Form, the Participant Information Sheet, a time schedule, and the Teams [108] link to the session.

The sessions with the experts were very similar to the ones conducted on the formative evaluation. After agreeing to being recorded, I sent them the PDF file with the script and they began executing the tasks. After they finished every task I asked them questions about the different features. After all the tasks were finished, I asked them their opinions on Wanda's potential. After the session finished, I sent them a questionnaire that included the SUS [16] and some questions about Wanda's potential.

The focus groups worked with similar mechanics, but I made sure to encourage discussion between all the participants. After they finished each of the tasks, I asked them explicitly about certain features, and when I saw that there were divided opinions, I asked them to comment on the thoughts of the other participants.

After each evaluation, I moved the video recordings from OneDrive to SharePoint, to avoid the risk of data loss due to OneDrive's retention policy of two months at the time of using the software. Once the videos were safely stored, I extracted the transcript of the videos to be able to analyse them faster.

F.5 Data analysis

In the figures F.1 and F.2, I present the code maps that I used for analysing the data of the requirement gathering phase.

Each of the themes identified had different frequencies throughout the evaluation. In the figure F.3 I present the hierarchy of the themes identified on the interviews conducted with the experts. Similarly, in the figure F.4 I show the same information but for the participants' focus groups.

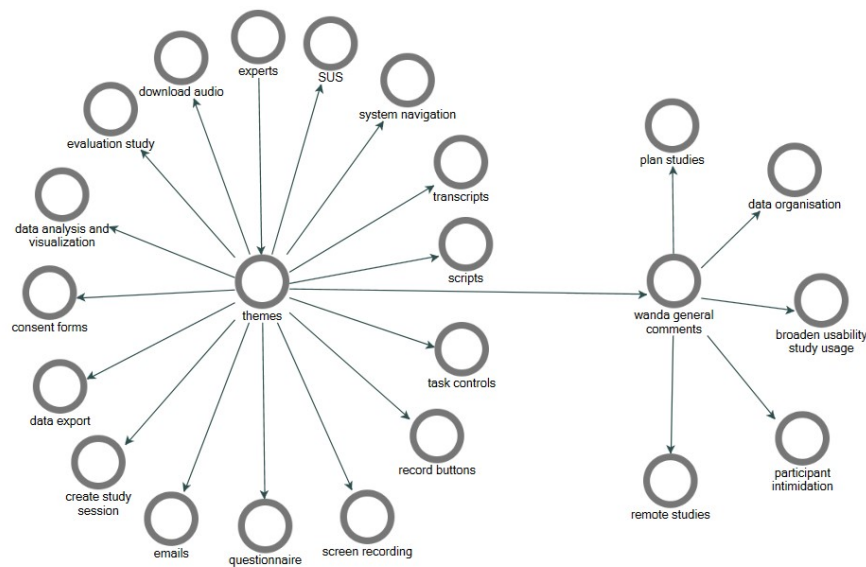


Figure F.1: Summative evaluation themes - experts

F.6 Results

F.6.1 Participant interface

Throughout the three focus groups conducted, all the participants successfully completed the evaluation of the mock e-commerce platform set up for them. It is important to note that all the evaluations were stand-alone, which means they did not require an expert to be present in the same session. Additionally, in the three focus groups, the system had to handle multiple requests simultaneously because they were all conducting the session in parallel, which also helped to identify the reliability of Wanda's backend infrastructure.

They began by logging in to the system, and overall their comments on the authentication flow were positive. They all agreed that the e-mail used for signing in was straightforward, and none of the users reported the e-mail going to their spam folder.

Once they began the evaluation and finished the first task, they were queried about the tutorial system described in section 7.1. All the participants expressed that the tutorial was clear and allowed them to understand the system's mechanics. However, two participants explained that they were confused by the fact that they had to mark the task as finished, as they would have expected the system to detect it automatically.

"I liked the tutorial pop up that had numbers in it. If can view one out of four two out of four, three out of four for the different alerts, which I thought was useful because

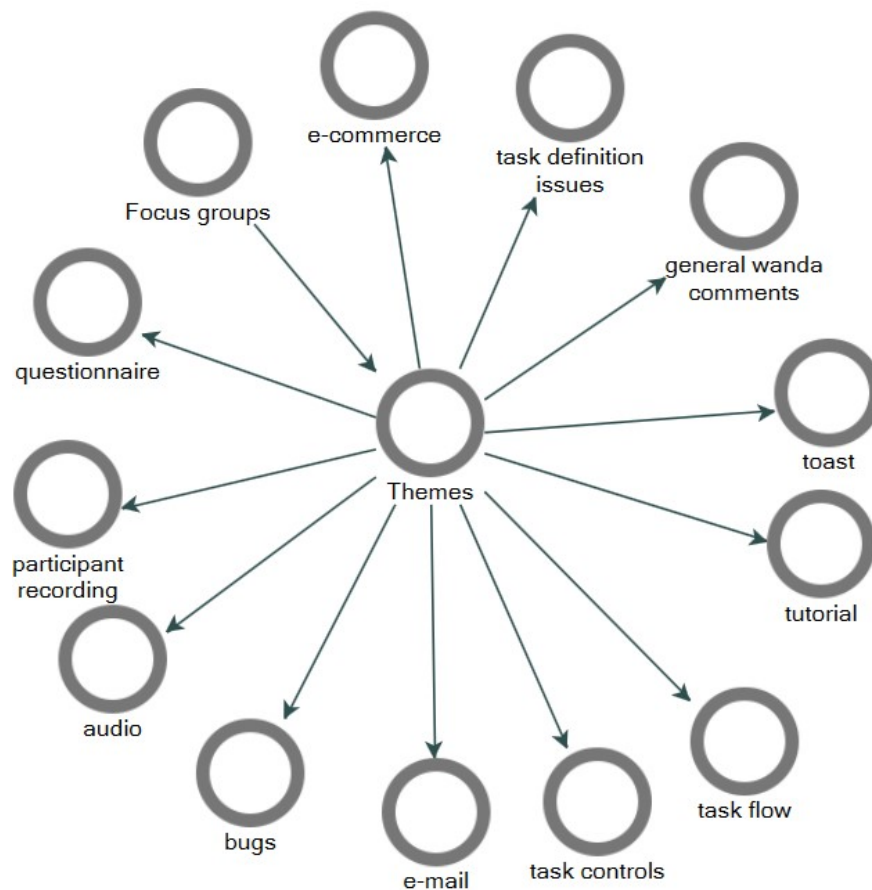


Figure F.2: Summative evaluation themes - participants

it gives you an idea of how many of these popups is going to be.” - P3.1.

As explained before, all the participants could complete the four tasks in the focus group. However, 2/14 did not provide their verbal thoughts while executing the task. In both cases, they explained that the main reason was that they did not understand that the system was recording their voice. This complaint was backed by 7/14 participants, who, despite having been able to speak their thoughts, they were not sure that the system was recording them. Overall, the recording system is the biggest complaint posed by the participants, as they expressed that Wanda is not clear enough when recording the participant’s voice.

“Also, one thing I wish that there was [...] is maybe some indication that my microphone is being used and like when it’s starting and when it’s when it’s stopped” - P3.4.

Participants were also asked about the task controls, and the flow for executing the

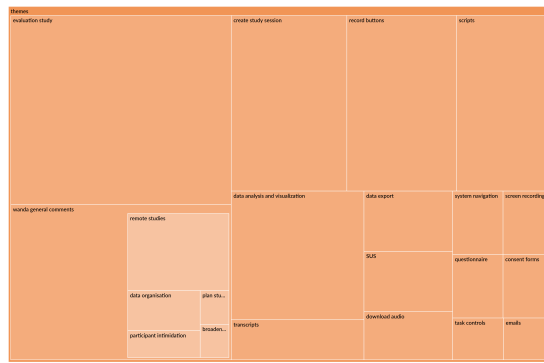


Figure F.3: Experts' summative evaluation theme hierarchy

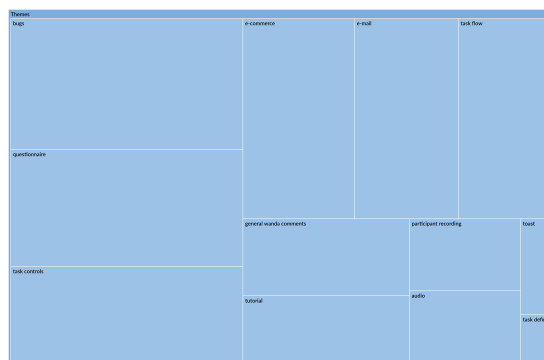


Figure F.4: Participants' summative evaluation theme hierarchy

task. Regarding the flow of the tasks, 12/14 of participants expressed that the flow was clear and easy to follow, primarily using adjectives like *clear*, *natural* or *easy*.

All the participants could mark tasks as finished and failed without significant issues, but 7/14 of participants expressed confusion about the buttons for marking a task as succeeded or failed. Since the interface provides a red cross icon for marking the task as failed, some participants confused it with a button for closing the task controls, even leading them to wrongly mark a task as failed.

While executing the evaluation of the e-commerce platform, 4/14 participants expressed technical issues with Wanda. One of those participants saw an error message expressing that “*There is a client application exception.*”. Upon further review, the participant said that they deactivated an extension for blocking cookies and that after doing that, the system worked as expected. I was not able to identify in the moment the issue with the other two participants, so I created new study sessions for them, and they successfully completed the evaluation.

The two participants with unknown technical difficulties were participants from the first focus group. After the session, I investigated the technical logs of the platform

and noticed that the database was running out of available connections, an issue well documented in Serverless deployments [64]. To avoid the problem in the following focus groups, I implemented two solutions: First, I set up the Prisma Data Platform [95], which handles database connections automatically from a Prisma schema. Second, I increased the memory and CPU on the database on AWS's [9] RDS service. Those two changes were successful, as the participants expressed no technical issues related to the backend in the following two focus groups.

8/14 participants also commented on the task and script recordings and the feature that allowed them to listen to the expert's voice. Three participants said that the audios did not work, and when further queried about their systems, they expressed that they were using the Safari [7] browser. Of the other participants, two said that they found the feature useful, and two commented that they thought the functionality would be helpful for users with visual disabilities.

"I mean, it is very inclusive and I think that's a very good feature to to bear in mind for. People obviously who might have an issue with reading. Because I know some of my older family members have a hard time reading website content in general because it's too small for them." - P3.4.

Participants were asked to fill out a questionnaire about the potential impact of Wanda, and their answers can be seen in figure F.5. The results indicate that participants mostly feel that Wanda allows them to understand how the Think Aloud [57] method works and that they can complete a session without an expert. However, data also suggest that participants believe that Wanda's most significant impact relies on making them feel confident when providing feedback about the site they are evaluating.

What are your thoughts regarding Wanda's potential impact

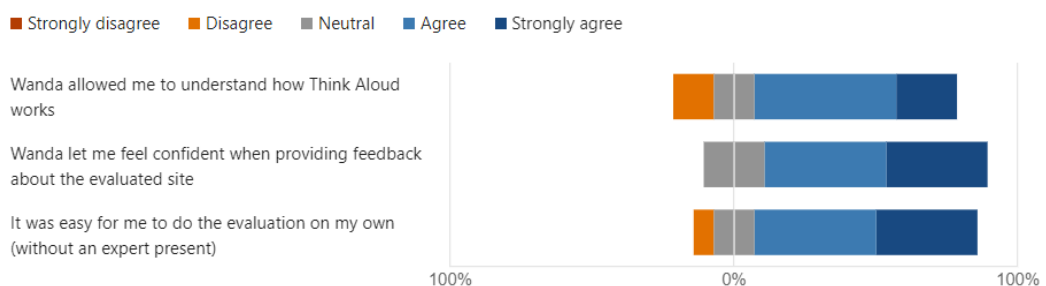


Figure F.5: Participants' opinions on Wanda's potential impact.

At the end of the evaluation, all the participants in the three focus groups were asked

to fill the System Usability Scale for Wanda. The results are presented in figures F.6 and annex F.7.1, showing that participants gave an average score of 81 to Wanda, which poses it on the *acceptable* range.

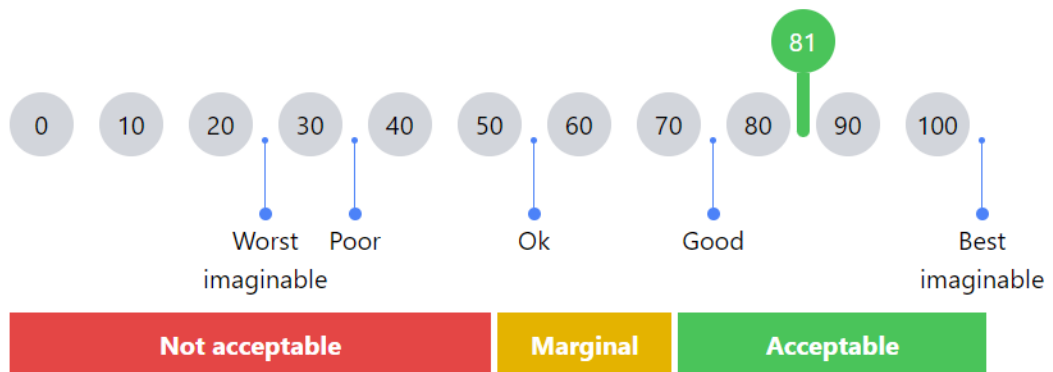


Figure F.6: SUS Results for participants' summative evaluation. Self-drafted image based on [126]

F.6.2 Expert interface

Overall, experts' comments about the tool were quite positive. All of them were able to set up an evaluation study for mock e-commerce platform, and they successfully could invite a participant and then analysed the results after the participant finished.

They all had positive comments regarding data export and visualisation, which was the most commented feature. 3/5 experts also commented that the automatic transcripts feature could help speed up evaluation analysis. Similarly, 3/5 commented about the charts on the evaluation results page, saying that having those visualisations along with the data export helps build their own reports.

“I think, doing processing or coding text is much easier than an audio. It's really nice that there is an automatic transcription like feature.” - P3.

All the experts commented on the System Usability Scale [16] visualisation. They all agreed that it is a good way of quickly knowing the level of usability of the system. However, one expert pointed out that it would be better to provide more insights on what *not acceptable*, *marginal* and *acceptable* mean.

“I like this scale, and the fact that you can also see 'very good' or between 'good' and 'best imaginable'. So you can interpret the usability fast.” - P3.

Despite the general comments being positive, 3/5 had complaints on the audio recording feature. They agreed that the system is not clear enough at providing the *recording* status, meaning that users can not easily identify that they are being recorded. This complaint is aligned to what some participants felt as reported on section F.6.1.

When queried about Wanda's potential impact regarding participants' feelings, 2/5 experts commented that Wanda does not necessarily allows participants to feel less intimidated. They explained that participants still need to be in front of a screen speaking, so there is still a barrier between the participant and their full potential of expressing their thoughts. 1/5 expert clarified, however, that by using Wanda, participants would be able to express their thoughts more easily.

3/5 experts commented on Wanda's impact on conducting remote usability studies. All of them used the adjective *straightforward* for referring to the process of creating, conducting and analysing a usability study, which indicates that the tool indeed has potential to aid in maintaining their studies more efficiently.

One expert commented that they felt that by being open-source, Wanda could help broaden the adoption of usability studies, since it can be more affordable for small companies.

"Yeah, I think it would be really easy for small companies to check their usability, because when you're in a small company you don't have a budget, so I think for startups this would be a nice tool for them to just improve their product." - P3.

One expert also explained that there was more work to do in terms of explaining the different methods allowed by the tool and how they differ, but they expressed that the tool might also be useful for students who may not be experts in usability but still need to conduct their studies, because it guides the user through all the process of designing and conducting the study.

The summary of experts' opinions on Wanda's impact can be found in figure F.7 where one can interpret that experts mostly agree that the most considerable potential of the tool lays in its ability to analyse evaluation studies faster. As also can be seen, experts are divided on Wanda's potential for recruiting participants more quickly. In that matter, one expert commented that it would be good to import a list of e-mails when creating the study session, for creating more than one at the same time, which may help increase the participant recruiting efficiency.

Additionally, experts mostly agree that Wanda may help participants feel less

intimidated, help them plan their studies faster and help them keep their studies more organised.

11. What are your thoughts regarding Wanda's potential impact (0 punto)

[Más detalles](#)

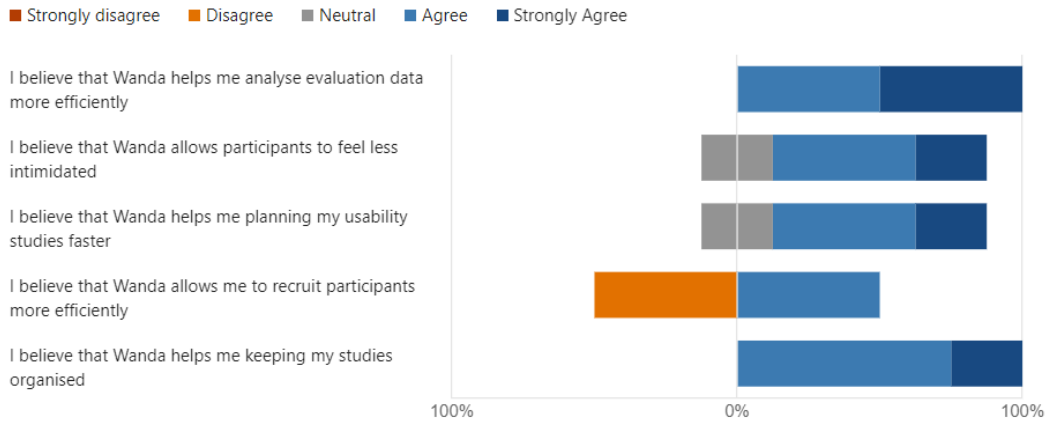


Figure F.7: Experts' opinions on Wanda's impact.

At the end of the evaluation, all the experts were asked to fill the System Usability Scale for Wanda. The results are presented in figures F.8 and annex F.7.2, showing that experts gave an average score of 94 to Wanda, which poses it on the *acceptable* range.

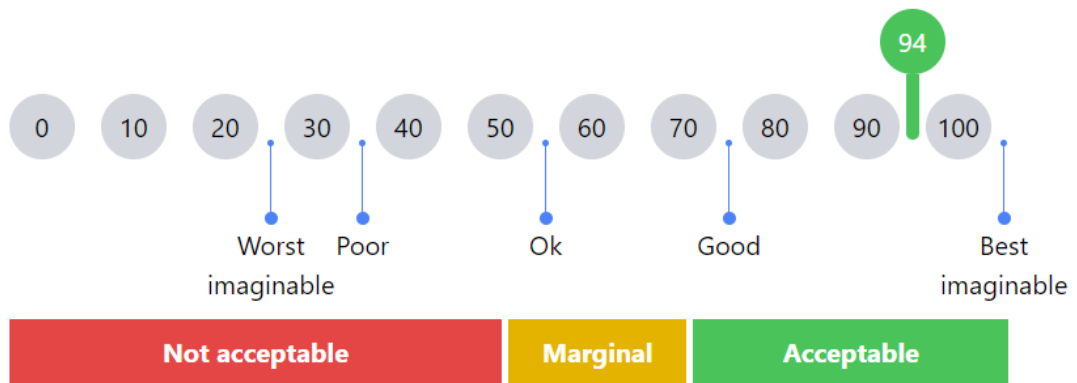


Figure F.8: SUS Results for experts' formative evaluation. Self-drafted image based on [126]

	I found the system unnecessarily complex	I think that this system makes the process of creating a Think Aloud session easy	I think that I would need support from a technical person every time I use this system	I thought the system was easy to use	I thought there was too much inconsistency in this system	I found the various functions in this system were well integrated	I found the system very cumbersome to use	I would imagine that most people would learn to use this system very quickly	I needed to learn a lot of things before I could get going with this system	I felt very confident using the system	SUS
P1	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Disagree	Agree	Strongly disagree	Agree	Strongly disagree	Agree	90
P2	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	Disagree	Agree	75
P3	Neutral	Agree	Disagree	Agree	Disagree	Neutral	Disagree	Neutral	Agree	Neutral	60
P4	Strongly disagree	Strongly agree	Disagree	Agree	Strongly disagree	Strongly agree	Disagree	Strongly agree	Disagree	Strongly agree	90
P5	Disagree	Neutral	Strongly disagree	Agree	Disagree	Agree	Disagree	Strongly agree	Strongly disagree	Agree	80
P6	Strongly disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	95
P7	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	97.5
P8	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	100
P9	Neutral	Agree	Disagree	Agree	Disagree	Strongly disagree	Disagree	Neutral	Disagree	Agree	62.5
P10	Disagree	Neutral	Disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	87.5
P11	Neutral	Neutral	Agree	Neutral	Disagree	Neutral	Neutral	Neutral	Disagree	Agree	55
P12	Strongly disagree	Strongly agree	Disagree	Agree	Disagree	Agree	Disagree	Strongly agree	Strongly disagree	Strongly agree	87.5
P13	Disagree	Agree	Neutral	Strongly agree	Disagree	Agree	Disagree	Strongly agree	Strongly disagree	Agree	80
P14	Disagree	Agree	Neutral	Agree	Disagree	Neutral	Strongly disagree	Strongly agree	Disagree	Agree	75

Figure F.9: SUS Results for participants' formative evaluation.

F.7 System Usability scale results

F.7.1 Participants

F.7.2 Experts

	I found the system unnecessarily complex	I think that this system makes the process of creating a Think Aloud session easy	I think that I would need support from a technical person every time I use this system	I thought the system was easy to use	I thought there was too much inconsistency in this system	I found the various functions in this system were well integrated	I found the system very cumbersome to use	I would imagine that most people would learn to use this system very quickly	I needed to learn a lot of things before I could get going with this system	I felt very confident using the system	SUS
P1	Strongly disagree	Agree	Strongly disagree	Agree	Strongly disagree	Agree	Disagree	Strongly agree	Strongly disagree	Agree	87.5
P2	Strongly disagree	Strongly agree	Strongly disagree	Agree	Strongly disagree	Agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	92.5
P3	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	Strongly disagree	Agree	Strongly disagree	Agree	92.5
P4	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Agree	Strongly disagree	Strongly agree	97.5
P5	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	Strongly disagree	Strongly agree	100

Figure F.10: SUS Results for experts' formative evaluation.

F.8 Future feature requirements for Wanda

- Add new evaluation methods such as cognitive walkthroughs [124].

Appendix G

Participants' information sheet

G.1 Requirement gathering

Participant Information Sheet for Initial Requirement Gathering

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator:	Cristina Adriana Alexandru
Researcher collecting data:	Daniel Saldarriaga (Main Researcher)
Funder (if applicable):	No

This study was certified according to the Informatics Research Ethics Process, RT number 2019/70801. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The study researchers are Daniel Saldarriaga, a postgraduate student in the University of Edinburgh School of Informatics, and Cristina Adriana Alexandru, his supervisor. This study is conducted as part of the postgraduate project of Daniel Saldarriaga.

What is the purpose of the study?

This project aims to develop an open-source online tool for helping experts conduct usability evaluation studies with users. The tool will allow the execution of both Think Aloud and Question-Asking protocol methodologies. This study aims to gather the initial requirements such a tool should fulfil. The results of this study will help us come up with a design of the tool that can potentially ease your future usability evaluation studies.

Why have I been asked to take part?

The reason why you are invited to participate in this study is because you are an expert in HCI and usability evaluation. We hope that you can use your previous valuable experience to suggest requirements for the tool.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. After this point, personal data will be deleted



and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI who is Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to participate in the study, we will organise an online one-to-one meeting with you over Zoom. Daniel Saldarriaga will lead the session, and the session will be audio and video recorded with your permission. During the meeting, you will be given a series of questions related to Think Aloud and Question-Asking Protocol methodologies. We will ask you a few questions about your experience, opinions, and suggestions regarding a future tool for supporting evaluation studies with users using both methods. In the end, we will show you initial sketches of a potential design for the tool, and we will ask for your comments regarding the idea. We will use your feedback to extract requirements and refine our design. The whole process will take around 30 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation. Your comments and answers will remain strictly confidential. Nothing you say will have any negative effect on your employment, appraisal, pay, degree, or anything else related to your working/study conditions.

Are there any benefits associated with taking part?

Although there are no physical benefits after this study, we do hope that the implementation of our tool will help you and your colleagues with the execution of usability evaluation with users.

What will happen to the results of this study?

The results of this study will be summarised in Daniel Saldarriaga's MSc dissertation. Moreover, they may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymised: We will remove any information that



could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymisation.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher Daniel Saldarriaga (s2092683@ed.ac.uk) and his supervisor Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk).

All electronic data will be stored on the School of Informatics' secure file servers. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the Principal Investigator: Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk)

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be emailed to you by Daniel Saldarriaga (s2092683@ed.ac.uk)

Alternative formats.



To request this document in an alternative format, such as large print or on coloured paper, please contact Daniel Saldarriaga (s2092683@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research



G.2 Formative evaluation

G.2.1 Experts

Participant Information Sheet for Formative Evaluation- Experts

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator:	Cristina Adriana Alexandru
Researcher collecting data:	Daniel Saldarriaga (Main Researcher)
Funder (if applicable):	No

This study was certified according to the Informatics Research Ethics Process, RT number 2019/70801. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The researchers of the study are Daniel Saldarriaga, who is a postgraduate student in the University of Edinburgh School of Informatics, and Cristina Adriana Alexandru who is his supervisor. This study is conducted as part of the postgraduate project of Daniel Saldarriaga.

What is the purpose of the study?

We are currently developing an open-source online tool for helping experts conduct usability evaluation studies with users. The study aims to evaluate our implementation formatively. The purpose is to find out the usability and the potential impact of the tool. This will help us improve the implementation. Hopefully we can improve this tool to make it useful for experts and users in the field of usability evaluation.

Why have I been asked to take part?

The reason why you are invited to participate in this study is because you are an expert in HCI and usability evaluation. We hope that you can use your previous valuable experience to suggest improvements to our implementation of the tool.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. After this point, personal data will be deleted



and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI who is Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to participate in the study, we will organise an online one-to-one meeting with you over Zoom. Daniel Saldarriaga will lead the meeting, which will be audio and video recorded with your permission. During this process you will be given a series of tasks for setting up, executing, and reviewing a Usability Evaluation study with users. You will be asked to report on your progress with each task. Then, we will ask you a few questions about your experience, opinions, and suggestions. At the end, you will be given a questionnaire through Microsoft Forms to survey your feelings about the potential impact of this prototype on your future work. The questionnaire will also contain the questions of the System Usability Scale (SUS), which are about your general views on the system's usability. The whole process will take around 30 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation. Your comments and answers will remain strictly confidential. Nothing you say will have any negative effect on your employment, appraisal, pay, degree, or anything else related to your working/study conditions.

Are there any benefits associated with taking part?

Although there are no physical benefits after this study, we do hope that the implementation of our tool will help you and your colleagues with the execution of usability evaluation with users.

What will happen to the results of this study?

The results of this study will be summarised in the Daniel Saldarriaga's MSc dissertation. Moreover, they may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any



information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher Daniel Saldarriaga (s2092683@ed.ac.uk) and his supervisor Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk).

All electronic data will be stored on the School of Informatics' secure file servers. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the Principal Investigator: Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk)

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be emailed to you by Daniel Saldarriaga (s2092683@ed.ac.uk)

Alternative formats.



To request this document in an alternative format, such as large print or on coloured paper, please contact Daniel Saldarriaga (s2092683@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research



G.2.2 Participants

Participant Information Sheet for Formative Evaluation- Users

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator:	Cristina Adriana Alexandru
Researcher collecting data:	Daniel Saldarriaga (Main Researcher)
Funder (if applicable):	No

This study was certified according to the Informatics Research Ethics Process, RT number 2019/70801. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The researchers of the study are Daniel Saldarriaga, who is a postgraduate student in the University of Edinburgh School of Informatics, and Cristina Adriana Alexandru who is his supervisor. This study is conducted as part of the postgraduate project of Daniel Saldarriaga.

What is the purpose of the study?

We are currently implementing an open-source online tool for helping experts conduct usability evaluation studies with users. The study aims to evaluate our implementation formatively. The purpose is to find out the usability and the potential impact of the tool. This will help us improve the implementation. Hopefully we can improve this tool to make it useful experts and users in the field of usability evaluation.

Why have I been asked to take part?

The reason why you are invited to participate in this study is because of your previous experience in participating to usability evaluation. We hope that you can use your previous valuable experience to suggest improvements to our implementation of the tool.

Do I have to take part?



No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI who is Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to participate in the study, we will organise an online group meeting with you over Zoom. Daniel Saldarriaga will lead the meeting, which will be audio and video recorded with your permission. During this process you will be given a series of tasks inside a website. You will be asked to use the tool for reporting your comments and asking questions about the site you are evaluating. You will be asked to report on your progress with each task inside the tool we will provide. Then, we will ask you a few questions about your experience, opinions, and suggestions, through a focus group. At the end, you will be given a questionnaire through Microsoft Forms to survey your feelings about the your experience while using the tool (and not about the evaluated the website). The questionnaire will also contain the questions of the System Usability Scale (SUS), which are about your general views on the system's usability. The whole process will take around 30 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation. Your comments and answers will remain strictly confidential. Nothing you say will have any negative effect on your working/study conditions.

Are there any benefits associated with taking part?

Although there are no physical benefits after this study, we do hope that the implementation of our tool will help you and your colleagues with the execution of usability evaluation with users.

What will happen to the results of this study?



The results of this study will be summarised in the Daniel Saldarriaga's MSc dissertation. Moreover, they may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher Daniel Saldarriaga (s2092683@ed.ac.uk) and his supervisor Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk).

All electronic data will be stored on the School of Informatics' secure file servers. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the Principal Investigator: Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk)

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.



If the research project changes in any way, an updated Participant Information Sheet will be emailed to you by Daniel Saldarriaga (s2092683@ed.ac.uk)

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Daniel Saldarriaga (s2092683@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research



G.3 Summative evaluation

G.3.1 Experts

Participant Information Sheet for Summative Evaluation- Experts

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator:	Cristina Adriana Alexandru
Researcher collecting data:	Daniel Saldarriaga (Main Researcher)
Funder (if applicable):	No

This study was certified according to the Informatics Research Ethics Process, RT number **349375**. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The researchers of the study are Daniel Saldarriaga, who is a postgraduate student in the University of Edinburgh School of Informatics, and Cristina Adriana Alexandru who is his supervisor. This study is conducted as part of the postgraduate project of Daniel Saldarriaga.

What is the purpose of the study?

We are currently developing an open-source online tool for helping experts conduct usability evaluation studies with users. The study aims to evaluate our implementation summatively. The purpose is to find out the usability and the potential impact of the tool. This will help us improve the implementation. Hopefully we can improve this tool to make it useful for experts and users in the field of usability evaluation.

Why have I been asked to take part?

The reason why you are invited to participate in this study is because you are an expert in HCI and usability evaluation. We hope that you can use your previous valuable experience to suggest improvements to our implementation of the tool.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. After this point, personal data will be deleted



and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI who is Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to participate in the study, we will organise an online one-to-one meeting with you over Teams. Daniel Saldarriaga will lead the meeting, which will be audio and video recorded with your permission. During this process you will be given a series of tasks for setting up, executing, and reviewing a Usability Evaluation study with users. You will be asked to report on your progress with each task. Then, we will ask you a few questions about your experience, opinions, and suggestions. At the end, you will be given a questionnaire through Microsoft Forms to survey your feelings about the potential impact of this prototype on your future work. The questionnaire will also contain the questions of the System Usability Scale (SUS), which are about your general views on the system's usability. The whole process will take around 45 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation. Your comments and answers will remain strictly confidential. Nothing you say will have any negative effect on your employment, appraisal, pay, degree, or anything else related to your working/study conditions.

Are there any benefits associated with taking part?

Although there are no physical benefits after this study, we do hope that the implementation of our tool will help you and your colleagues with the execution of usability evaluation with users.

What will happen to the results of this study?

The results of this study will be summarised in the Daniel Saldarriaga's MSc dissertation. Moreover, they may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any



information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher Daniel Saldarriaga (s2092683@ed.ac.uk) and his supervisor Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk).

All electronic data will be stored on the School of Informatics' secure file servers. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the Principal Investigator: Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk)

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be emailed to you by Daniel Saldarriaga (s2092683@ed.ac.uk)

Alternative formats.



To request this document in an alternative format, such as large print or on coloured paper, please contact Daniel Saldarriaga (s2092683@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research



G.3.2 Participants

Participant Information Sheet for Formative Evaluation- Users

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator:	Cristina Adriana Alexandru
Researcher collecting data:	Daniel Saldarriaga (Main Researcher)
Funder (if applicable):	No

This study was certified according to the Informatics Research Ethics Process, RT number 2019/70801. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The researchers of the study are Daniel Saldarriaga, who is a postgraduate student in the University of Edinburgh School of Informatics, and Cristina Adriana Alexandru who is his supervisor. This study is conducted as part of the postgraduate project of Daniel Saldarriaga.

What is the purpose of the study?

We are currently implementing an open-source online tool for helping experts conduct usability evaluation studies with users. The study aims to evaluate our implementation formatively. The purpose is to find out the usability and the potential impact of the tool. This will help us improve the implementation. Hopefully we can improve this tool to make it useful experts and users in the field of usability evaluation.

Why have I been asked to take part?

The reason why you are invited to participate in this study is because of your previous experience in participating to usability evaluation. We hope that you can use your previous valuable experience to suggest improvements to our implementation of the tool.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI who is Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk). We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to participate in the study, we will organise an online group meeting with you over Zoom. Daniel Saldarriaga will lead the meeting, which will be audio and video recorded with your permission. During this process you will be given a series of tasks inside a website. You will be asked to use the tool for reporting your comments and asking questions about the site you are evaluating. You will be asked to report on your progress with each task inside the tool we will provide. Then, we will ask you a few questions about your experience, opinions, and suggestions, through a focus group. At the end, you will be given a questionnaire through Microsoft Forms to survey your feelings about the your experience while using the tool (and not about the evaluated the website). The questionnaire will also contain the questions of the System Usability Scale (SUS), which are about your general views on the system's usability. The whole process will take around 30 minutes.

Are there any risks associated with taking part?

There are no significant risks associated with participation. Your comments and answers will remain strictly confidential. Nothing you say will have any negative effect on your working/study conditions.

Are there any benefits associated with taking part?

Although there are no physical benefits after this study, we do hope that the implementation of our tool will help you and your colleagues with the execution of usability evaluation with users.

What will happen to the results of this study?



The results of this study will be summarised in the Daniel Saldarriaga's MSc dissertation. Moreover, they may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 2 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher Daniel Saldarriaga (s2092683@ed.ac.uk) and his supervisor Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk).

All electronic data will be stored on the School of Informatics' secure file servers. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the Principal Investigator: Cristina Adriana Alexandru (Cristina.Alexandru@ed.ac.uk)

If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.



If the research project changes in any way, an updated Participant Information Sheet will be emailed to you by Daniel Saldarriaga (s2092683@ed.ac.uk)

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Daniel Saldarriaga (s2092683@ed.ac.uk)

General information.

For general information about how we use your data, go to: edin.ac/privacy-research



Appendix H

Participants' consent form

H.1 Requirement gathering

Participant number:___ **2022/61691** ___

Participant Consent Form

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator (PI):	Cristina Alexandru
Researcher:	Daniel Saldarriaga (s2092683@ed.ac.uk)
PI contact details:	Cristina.Alexandru@ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1. I agree to being audio recorded.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

2. I agree to being video recorded.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

3. I allow my data to be used in future ethically approved research.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

4. I agree to take part in this study.

<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------

Yes No

Name of person giving consent

Date
dd/mm/yy

Signature

Name of person taking consent

Date
dd/mm/yy

Signature



H.2 Formative evaluation

Participant Consent Form

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator (PI):	Cristina Alexandru
Researcher:	Daniel Saldarriaga (s2092683@ed.ac.uk)
PI contact details:	Cristina.Alexandru@ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1. I agree to being audio recorded.

Yes	No

2. I agree to being video recorded.

Yes	No

3. I allow my data to be used in future ethically approved research.

Yes	No

4. I agree to take part in this study.

Yes	No

Name of person giving consent

Date
dd/mm/yy

Signature

Name of person taking consent

Date
dd/mm/yy

Signature



H.3 Summative evaluation

Participant number: _____ **349375** _____

Participant Consent Form

Project title:	Towards a Tool to Support Think Aloud and Question-Asking Protocol Evaluation with Users
Principal investigator (PI):	Cristina Alexandru
Researcher:	Daniel Saldarriaga (s2092683@ed.ac.uk)
PI contact details:	Cristina.Alexandru@ed.ac.uk

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1. I agree to being audio recorded.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

2. I agree to being video recorded.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

3. I allow my data to be used in future ethically approved research.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

4. I agree to take part in this study.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

Name of person giving consent

Date
dd/mm/yy

Signature

Name of person taking consent

Date
dd/mm/yy

Signature

