UNIVERSITY OF TWENTE

Combining think-aloud protocols with eye tracking technology for usability: An exploratory comparative analysis

A Master Thesis

Albert R. Berkhoff, s1569511

6 August, 2020

Author Note
Albert R. Berkhoff, Faculty of Behavioural, Management and Social sciences, University of Twente.
Supervised by: Dr. S. Borsci & P. Slijkhuis, MSc.

Abstract

The present study examined different think-aloud protocols with eye tracking technology in usability testing in an exploratory and comparative way. A focused selection was made from a list of usability methods and usability techniques to deepen the understanding of think-aloud protocols and eye tracking technology. The concurrent and retrospective think-aloud protocols have been used together with a gaze-measuring (classic) eye tracker and an (cued) eye tracker with a vision bubble to create four conditions. To compare the four conditions, a review from previous research with similar usability tests had been done. From the review, six criteria have been selected by which the four conditions can be compared. The six criteria have been used with the questionnaire UMUX-lite and the instrument Rating Scale Mental Effort to do the comparative study. The results of the comparative study suggest that the retrospective think-aloud protocol which utilizes the classic eye tracker fits the criteria better than the three other conditions. Furthermore, the results suggest that concurrent and retrospective think-aloud protocols which use the cued eye tracker fit the criterion regarding the participants' experience better than the two think-aloud protocols which use the classic eye tracker. The results also suggest that the two think-aloud protocols which use the cued eye tracker fit the criteria regarding the questionnaire UMUX-lite and the time spent on a task worse than the two think-aloud protocols which use the classic eye tracker. In conclusion, the present study can be seen as an in-depth exploration into the world of usability testing and encourages investigating think-aloud protocols and eye tracking further.

Keywords: Usability; think-aloud protocols; eye tracking; Rating Scale Mental Effort; UMUX-lite.

Contents

1. Introduction

Nowadays, businesses, institutions and companies can own and manage an online website to, for example, sell their products, but also to promote the business to those who are interested. This website often has a focus on a specified group of users. However, this group of users can differ in both personality and goals to achieve on the website. Therefore, the group of users can achieve different results and experiences from using the website. To measure the extent of this group of users that can achieve specified goals on the website, the measurement usability is used and tested. This is done through three aspects, namely the effectiveness, efficiency, and the satisfaction of the website (ISO, 2018). The effectiveness is the accuracy and completeness with which users achieve specified goals, while efficiency is the effort of achieving results of certain accuracy and completeness. Meanwhile, satisfaction is the extent to which the user's physical, cognitive, and emotional responses meet the user's needs and expectations. The physical, cognitive, and emotional responses from the users are a result from the usage of a product, system, or service, such as a website.

The measurement of these aspects of usability is an important factor for website administrators from business, institutions, and companies to increase the success of the website for several reasons. A study from Palmer (2002) has suggested that the success of B2C (Business to Customer) websites is strongly related to usability. This means that utilizing usability in order to improve a B2C website has increased the frequency of use, user satisfaction, and intent to return to the website. Another study has found that increasing the usability has a positive effect on the trustworthiness of a website according to its users (Roy, Dewit, & Aubert, 2001). In this study, trustworthiness is defined as the compound of the perceived ability of the website, the perceived benevolence of the website, and the perceived integrity of the website. This compound was measured through five factors of usability, namely the ease of navigation, consistency, ease of learning, perception, and support. These five factors are positively correlated towards the compound of the three aspects of trustworthiness. Another study has found similar results in the trustworthiness of websites, which means that usability has a positive relation with trust in a website from a user (Casaló, Flavián, & Guinalíu, 2007). Furthermore, this study has found that the level of trust within the users is positively related to the commitment from the users to the website.

Another purpose of testing for the measurement usability, also known as usability tests, is to capture the current user experience within the interaction between the user and the product, system, or service (Whiteside, Bennett, & Holtzblatt, 1988). This means that usability tests focus on the users of a product, system, or service and what their experiences are from using the product, system, or service. These experiences, also known as user experience or UX, can be defined as the perceptions and responses from a person that are results from the use and/or anticipated use of a product, system, or service (ISO, 2018). At the heart of user experience is usability, whereby usability can be used by a product, system, or service to positively alter the perceptions and responses from users (Hassan & Galal-Edeen, 2017). It is important to notice that products, systems, or services are dynamic and ever-changing, which has an effect that usability of a product, system, or service is constantly shifting and therefore never wholly complete. Therefore, usability testing captures the current state of affairs from a product, system, or service that is being tested. The focus of the tests is to capture the negative state of affairs and the positive state of affairs. The negative state of affairs is known as the usability problems, and it has as purpose to serve as a basis in improving the usability to increase the success of a product, system, or service. On the other hand, the positive state of affairs should be strived for in usability testing and has already a contribution to the success of a product, system, or service. Therefore it is important to focus on both the positive strengths and negative weaknesses.

1.1 Methods to Test for the Usability Problems

There are seven methods to test for usability problems and strengths according to Babich (2019). These methods are focused on the performances of the users, to capture the most authentic perceptions and responses from the users. The seven methods can be summarised as follows:

- Guerrilla testing: a method in which a website is tested by random participants that are collected in a public location such as a shopping mall. Often these random participants are given a small reward for their participation, such as a cup of coffee. This method is ideal in testing for a product that has a broad and mixed target group, because there is an increased chance that a mixed set of opinions is gathered. The time of guerrilla tests is limited and therefore should be as short as possible, since passer-by's often do not have a large amount of time.
- Lab testing: a method in which a website is tested on its usability in a laboratory. In these lab tests, researchers can go in-depth with the usability tests, due to the fact that the laboratory enables the researchers to use intensive techniques to investigate the reasoning behind participants' behaviour.

However, the laboratory can differ from the environment from the users of the final product. Therefore the results can be skewed and as a result the required changes of a website can no longer work. Researchers should take this in consideration.

- Unmoderated remote usability testing: a method in which participants can do the usability test at any place and any time where and when the participants desire. This method is cost-efficient and can be done by a multitude of participants at the same time. Therefore the sample of participants is large, but every result has the chance to be shallow. This means that more complicated questions can be left unanswered.
- Contextual inquiry: a method in which participants show for example their preferences with using websites in general. This method can be considered not so much as a usability test, but more like observational testing. The reason for that is, because the participants used in this test method are observed in their own environments without any interference from the researchers. This means that the participants are users of a product, system, or service that already have experience using the product, system, or service.
- Phone interview: a method in which participants are interviewed by using the phone while participants complete certain tasks. The benefit of this method is that participants can complete the usability tests in known environments for the participants from all over the world. That requires a researcher with exceptional communication skills to guide the participants through a phone connection, in order to make the interaction between researcher and participant as clear as possible.
- Card sorting method: a method which is mostly used for the navigation of a website. In the card sorting method, participants sort cards in a method that is logical to the participant. The cards often have terms that are used in the website that is being tested. An example is the navigation structure from an online web shop. This method explains to the researchers what logical navigation is according to the participants.
- Session recording: a method in which participants will do a certain task on a website and are recorded while working on the task. These recordings will first

be made anonymous and then be analysed. This method is often used in combination with the other mentioned methods to maximise the results.

In order to choose one or multiple methods, Hotjar (2019) suggested using two different criteria that can help decide with selecting one or multiple methods. The first criterion is whether the researchers want to have moderated or unmoderated usability testing. With moderated usability testing, participants are testing the target website while the researchers observe the participants and guide the participants where required. While with unmoderated usability testing, participants are left alone while testing the target website. In general, the reasoning and motivation behind certain behaviour from the participants is only observed with moderated usability testing, while unmoderated usability testing is economical more favourable and has a focus on behaviour patterns. The second criterion is whether the participants test the target website in-person or on a remote location. In-person testing happens when the participant completes the test while a researcher is physically present. And the remote testing happens when the participants complete the test without the supervision of a researcher, through for example the internet or a phone connection. The beneficial side of in-person testing is that the acquired data is more extensive in the way that body language and facial expression are included in the data, while the beneficial side of remote testing is that a larger target group is reached with using fewer resources. An overview of the methods with corresponding criteria for helping in the decision to select one or multiple methods can be seen in Table 1.

Table 1

Overview of Methods with Focus on Performance of Participants with Explanation and Corresponding Criteria, Gathered from Babich (2019) and Hotjar (2019)

Methods	Short description	Criterion: moderated or unmoderated?	Criterion: in-person or remote?
Guerrilla testing	Testing participants in random (crowded) locations such as a shopping mall	Moderated	In-person
Lab testing	Testing participant in a laboratory setting	Moderated	In-person
Unmoderated	Testing participants	Unmoderated	Remote

remote usability testing	unsupervised from familiar environments for the participants		
Contextual inquiry	Testing/observing users in their natural environments	Unmoderated	In-person
Phone interview	Testing remote participants through a phone connection	Moderated	Remote
Card sorting	Testing the participants by using cards that participants need to sort	Moderated	In-person
Session recording	Testing participants and recording the tests in order to analyse the recordings	Unmoderated	Remote

1.2 Techniques to Test for the Usability

Besides using the planning around the usability test that is known as methods, techniques should be used to test for the usability. A technique can be defined as the way of doing an activity that requires skill (Cambridge Dictionary, n.d.). In other words, a technique means how an activity such as a usability test can be done with a certain skill. Thus a technique in the context of usability tests focuses on how the usability test should be performed by both participant and researcher.

There is a multitude of techniques that can be used in usability tests. According to Usability Home (n.d.) and Poole and Ball (2005), there are ten different independent usability testing techniques. The ten techniques can be summarised as follows:

- Coaching technique: a technique in which a participant is testing a website while having an expert sitting next to the participant. This expert can answer any questions related to the product, system, or service that the participant can have during the testing phase. The justification for this technique is that it is used to discover the information needs of the users, so that training and documentation can be improve for the product, system, or service, as well as an improvement for the product, system, or service.

- Co-discovery learning: a technique which is used by two participants at the same time. The two participants can help each other in difficult times while testing, and the participants are encouraged to explain so that both participants can understand each other. It is preferable that the participants should know each other, to make the co-discovery as smooth and easy as possible.
- Performance measurement: a technique that focuses on what the quantitative performances of the participants are. It is preferable to use this technique in a laboratory set-up, so that the measurements are as accurate as possible. Examples of performance measurement are the time that a participant spent on the task, or whether a participant can or cannot complete a certain task.
- Question-asking protocol: a technique in which the researchers are prompting the participants by asking them relevant questions. These questions can help the researchers to get an insight in the mental model of the product, system, or service that is being tested from the participants. In this protocol, it is encouraged to ask both direct questions and more broad questions.
- Remote testing: a technique that can be combined with almost any other technique and with this technique, the usability tests are happening in separated places and/or times for participants and researchers. Usually computers or telephones are used to make a connection between the researcher and participant to perform the usability tests.
- Retrospective testing: in this technique, participants view a recording of their own performance and the participants provide comments on their performance. These comments will explain the motives of the participants' actions during the testing phase.
- Shadowing technique: in this technique, an expert user in the domain sits with the researcher and explains to the researcher what the participant is doing while testing a product, system, or service. This technique is appropriate when the participant cannot think-aloud during the test phase.
- Teaching technique: a technique is done by two participants. The first participant works with the product, system, or service to acquire some familiarity and experience with the product, system, or service by accomplishing tasks. After the first participant is done with gathering experiences and is ready for the next step, the second participant is introduced.

The second participant is a naïve and new user to the product, system, or service, and both participants together try to solve a set of tasks. However, the first experienced participant cannot actively solve the tasks.

- Think-aloud protocol: in this technique participants verbalise and think-aloud about their thoughts, opinions, and feelings while performing tasks on and working with a product, system, or service. This technique gives a direct insight into the mental model from the participants, but also the interaction between the participant and the product, system, or service. This technique can be considered important to the present study because the present study focuses on the performance of the users and participants of a product, system, or service, to capture the most authentic perceptions and responses from the users and participants.
- Eye tracking: a technique which can track the gaze of a participant by using a device that can shine infra-red light. This light is reflected in the eyes of its users, and therefore the gaze can be constantly monitored by the device. This technique does fit in the realm of usability tests, and can be used in combination with one or multiple other techniques. Eye tracking can give a more deepened understanding about the usability of a product, system, or service. A company that develops and manufactures eye tracking devices is Tobii (Tobii Group, 2020).

A large variety of methods and techniques can help to improve the usability of products, systems, and services. To know the strengths and weaknesses of these methods and techniques may help practitioners to define efficiently and effectively their usability testing setup. Different methods and techniques have different unknown advantages and disadvantages. Due to the many possible combinations of methods and techniques, it can be difficult or even problematic to grasp the full understanding of what the strengths and weaknesses are from the different possible setups, especially when eye tracking technology is also involved. The present work rationale is to compare and test different setups of usability testing supported by eye tracking devices, in order to create an understanding about the strength and weaknesses of different usability testing setups. Therefore to select different setups, the key elements of these setups are defined. Furthermore, the criteria that establish the possibility of comparing the setups are defined by means of a literature review.

2. Definition of Setups and Criteria

2.1 The Definition of Key Elements

An eye tracking usability test setup can be designed by combining at least the following five key elements:

- Environment: the allocation of an environment in which the usability test takes place has a direct influence on the selection of methods and techniques, and therefore is critical to be done as initial starting point (Babich, 2019). The environment of the test can be either inside the laboratory, or outside the laboratory. A laboratory has as advantage that it can provide room and possibility for more complicated and profound eye tracking devices, while doing usability tests outside a laboratory forces to use versatile but small and moveable eye tracking device that can be used virtually in any place or room.
- Types of eye tracking devices: there are three eye tracking devices developed and manufactured by Tobii Technology (Tobii Group, 2020). The first eye tracking device, which is a special designed computer with a built-in eye tracking system, is used as an assistive technology tool for communication (Tobii Dynavox, 2020). The second device is a bar which can be mounted underneath or at the bottom of a computer screen, to track the eyes of the user of the computer. This bar is non-obtrusive for its users, while still collecting reliable and relevant data. The last device is a pair of glasses or other wearable devices. This eye tracking device is more obtrusive than in comparison to the mounted bar, but the sensors in the wearable eye tracking device that measure the whole eye tracking are mere centimetres in distance in front of the eyes.
- Level of moderation: usability tests can either be moderated by for example a researcher or be unmoderated (Hotjar, 2019). A usability test supported by eye tracking is more likely to be moderated, because the first step with working with eye tracking devices is calibrating the eye tracking device onto the eyes of its user. Unless the user, which is in this case the participant in the usability test, owns such a device, the user requires assistance with the calibration. The researcher that helps with the calibration could step out to do an unmoderated usability test, or decide to moderate the usability test.
- Required protocol of verbalisation: there are two protocols of verbalisation, concurrent think-aloud protocol (CTAP) and retrospective think-aloud protocol

(RTAP). With CTAP, participants think aloud while working on a task. And with RTAP, participants think aloud after they are done with a task. RTAP lasts almost double in time in comparison to CTAP, but CTAP can suffer from reactivity from users (Van den Haak, De Jong, & Schellens, 2003). Reactivity works out in either a better performing participant as a result of a more structured working process, or a worse performing participant as a result of a double increased workload (Russo, Johnson, & Stephens, 1989).

- Measures and metrics: the eye tracking devices manufactured by Tobii Technology come with the analysis software Tobii Pro Lab (Tobii Pro Lab, 2020). This software can analyse the data generated by the eye tracking devices, such as areas on the video with gazes of the participants that are interesting for researcher to further analyse. This is also known as the areas of interest (AOI) feature.

The combination of these five key elements creates different possible setups. In line with literature, each setup is comparable to the others by six criteria: performance, the amount of usability problems, the severity level, the types of usability problems, the detection method, and the participants' experience.

2.2 Six Criteria to Compare Testing Setups

Appendix A provides an overview of the six criteria to compare setups previously applied in usability studies. The six criteria can be summarised as follows:

- Performance: the two aspects that are important in the criterion performance is the time that the participants on the task and whether participants finish the task successfully. In general, a usability test is considered easier if the time spent on it by the participants is lower and the rate of successful completion is higher. This can then be explained by either by the difficulty of the task, the difficulty of the technique, or the participants self. For this criterion, the technique 'Performance measurement' is used.
- The amount of usability problems: a usability testing setup can be considered more fruitful in the case of a larger amount of usability problems that is discovered by this setup in comparison with the other setups. This criterion is considered to be "The most common way" (Alhadreti & Mayhew, 2018).
- The severity level: there are four levels of how sever a usability problem can be. The first level is 'critical' and means that the problem prevents the user from completing

the task. The second level is ‘major’ and major problems create significant delay and frustration with its users. The third level is ‘minor’ and means that the problem has a minor effect on the usability. And the last level is ‘suggest’ and these problems means subtle and possible enhancements or suggestions of improvement.

- The types of usability problems: to summarise the studies that focus on this criterion, there are four different types of usability problems. The four types of usability problems focus on either the content of the webpage, the technical issues behind a website, the design that the website uses, and with how the navigation works on the website.
- The detection method: there are three ways how a usability problem is detected. The first way is that a problem is detected by the verbalisation of the participant; the second way is that a problem is detected by the observation of the researcher, and the last way is a combination of both previous ways.
- Participants’ experience: questionnaires with Likert scales can assess how the participants experienced the usability tests. These questionnaires focused on aspects such as the tiredness of the participants, the opinions of the participants about the research team that is present during the usability tests, and how time-consuming the usability test was for the participants.

3. Selection of Setups and Criteria for the Present Study

The comparison will be started by reducing the amount of methods and techniques that are discussed in previous sections by selecting a limited amount of methods and techniques. The reason for this is due to the limited time and resources the researchers have. Therefore the selection is based on the availability of time and resources from the researchers, and the five key elements of usability testing with eye tracking support presented earlier.

The first step is to allocate an environment in which the participants and researchers will use the methods and techniques. This is based on the key element of the allocation of an environment in which usability testing supported by eye tracking devices will take place. The research team of the current study is affiliated with the University of Twente, and therefore the team has access to a laboratory environment known as 'The BMS Lab'. This is a laboratory environment specifically for the faculty of Behavioural, Management and Social sciences, and consists of a multitude of different rooms in which laboratory experiments can be performed. The most suitable rooms for a usability testing setup with eye tracking possibility are the 'flexperiment' rooms (BMS Lab, n.d.).

With a fitting environment allocated, the methods that will be used in the current study are selected. This selection is made on the base of the two criteria from Hotjar (2019). To capture both criteria, two extremes is tried to be compared to each other. This means that two methods are selected, based on the key element of the level of moderation; one method with the criteria moderated and in-person, and the other method with the criteria unmoderated and remote. The method that is most suitable with the criteria moderated and in-person is lab testing since this method is moderated by researchers and participants must be present in a lab to do this method. The other method is an adaptation of the session recording method. Since a laboratory will be available for the present study to use the first method, the second method will also be recorded in the same laboratory. The reasoning behind this is to minimise differences when comparing both methods. The adaptation of the session recording testing method is that participants will be tested in-person in the laboratory instead of remote testing. Therefore the main thing that will be analysed in the methods is the criterion focus on the moderation.

Furthermore, the techniques that will be used in the current study are selected. The techniques are selected on the availability of resources and time for the researchers. Therefore, some techniques cannot be selected due to limited resources and time. But it is in the interest of the researchers to analyse as much techniques as possible, thus the techniques that will not be used are described now. The first two techniques that will not be used are the coaching technique and the shadowing technique. Both techniques require an expert that will help the participants in their usability journey. However, the present study is done by a research team in which one researcher will be responsible for the gathering of the data. This researcher is a master student at the University of Twente and therefore not (yet) an expert in the domain that will be tested on its usability. The next technique is the remote testing technique, and will not be used because a laboratory will be used; thus the need of remote testing is diminished. Besides this, the time is limited to collect all the data, and time will therefore be spend on testing in a laboratory setting. The next two techniques that will not be used are the co-discovery learning technique and the teaching technique. Both techniques will not be used because at least three people, namely two participants and one researcher, are needed to perform the techniques. Unfortunately, the maximum of persons allowed in the 'flexperiment' rooms are two persons (BMS Lab, n.d.) and thus the co-discovery learning technique and the teaching technique cannot be used in the current study. The rest of the techniques, namely the performance measurement, the question-asking protocol, the retrospective testing technique, the think-aloud protocol, and the eye tracking technique can and will be used in the current study.

The present study will focus on comparing the methods and techniques of eye tracking usability tests with each other. To perform this comparative analysis, the present study established five key elements of usability testing and six criteria through literature review. From the five key elements, two methods and five techniques of usability testing with eye tracking technology were selected to identify strengths and weaknesses of different setups for research. The six criteria are used in the comparative analysis for the two methods and five techniques. Furthermore, the present thesis is composed of two parts. An initial study was done test the setups and eventually compare the setups. However, an error in the allocation of participants during the randomisation resulted in an analysis that, while it could be considered a good usability analysis, could not be used to compare the setups in an efficient and effective manner. Due to this mistake, the initial study was treated as a pilot. The second study adjusted the procedure in order to assign participants correctly to the different setups for the testing by also enabling comparative analysis.

4. The Exploratory Pilot

4.1 Method of the Pilot

The two methods and five techniques selected in the previous section are at the base of the present study. The techniques regarding eye tracking, retrospective testing, and think-aloud protocol formed the conditions that will be used in the experimental phase of the present research. The think-aloud protocol technique can be distinguished in two different ways to implement the protocols. The two ways to implement think-aloud protocol is known as concurrent think-aloud protocol (CTAP) and the retrospective think-aloud protocol (RTAP). In the CTAP, participants work on one or multiple tasks and express their thoughts, feelings, and opinions by thinking out loud about what the participants are working on and working with. This thinking aloud and working on the tasks are happening at the same time with the CTAP. The difference with RTAP is that the moment that participants have to express their thoughts, feelings, and opinions by thinking aloud is after the participants are done with the tasks. That means that participants will work on the tasks in silence, and after the tasks are done participants will verbalise what their thoughts, feelings, and opinions are. The RTAP also has roots in the retrospective testing technique. Often this retrospective verbalisation is done with the guidance of video or audio footage from the performance of the participant.

This study was designed as a 2X2 design, in which the CTAP and RTAP is tested with support from eye tracking technology. A distinction can be made into two conditions in which the two protocols will be tested, and are known as the classic conditions and the cued conditions. All conditions are involved with eye tracker technology, namely the eye tracking device Tobii Pro X3-120 for the classic conditions and the eye tracking device Tobii 4C for the cued conditions. The difference between eye tracking devices will be further explained in the materials section. The difference of the classic and cued conditions is in the feedback that the participants will receive during or after the tasks. In the classic condition, participants will receive no additional feedback cues other than what is deemed to be regular to the concurrent and retrospective think-aloud protocols. In the cued condition, participants are receiving cues on where their gaze is during the tasks. These cues are visualised on the screen of the participant by a vision bubble that simulates the gaze of the participant, as can be seen in Figure 1. This means for the concurrent think-aloud protocol that participants are receiving additional cues by only the vision bubble, while for the retrospective think-aloud protocol participants receive the additional cues as vision bubble and a playback video. The two conditions and two protocols give this study a total of four conditions, which can be found in Table 2. Added in this table are also the cues from each condition.

Table 2
the 2X2 Design from the Current Study Processed into an Overview, with the Two Different Think-Aloud Protocols and the Classic and Cued Condition. Added in Table are the Cues Participants will Receive during the Experiment.

	Classic Condition	Cued Condition
Concurrent Think-Aloud Protocol	Classic CTAP →No cues	Cued CTAP →Vision bubble
Retrospective Think-Aloud Protocol	Classic RTAP →Video	Cued RTAP →Video and vision bubble

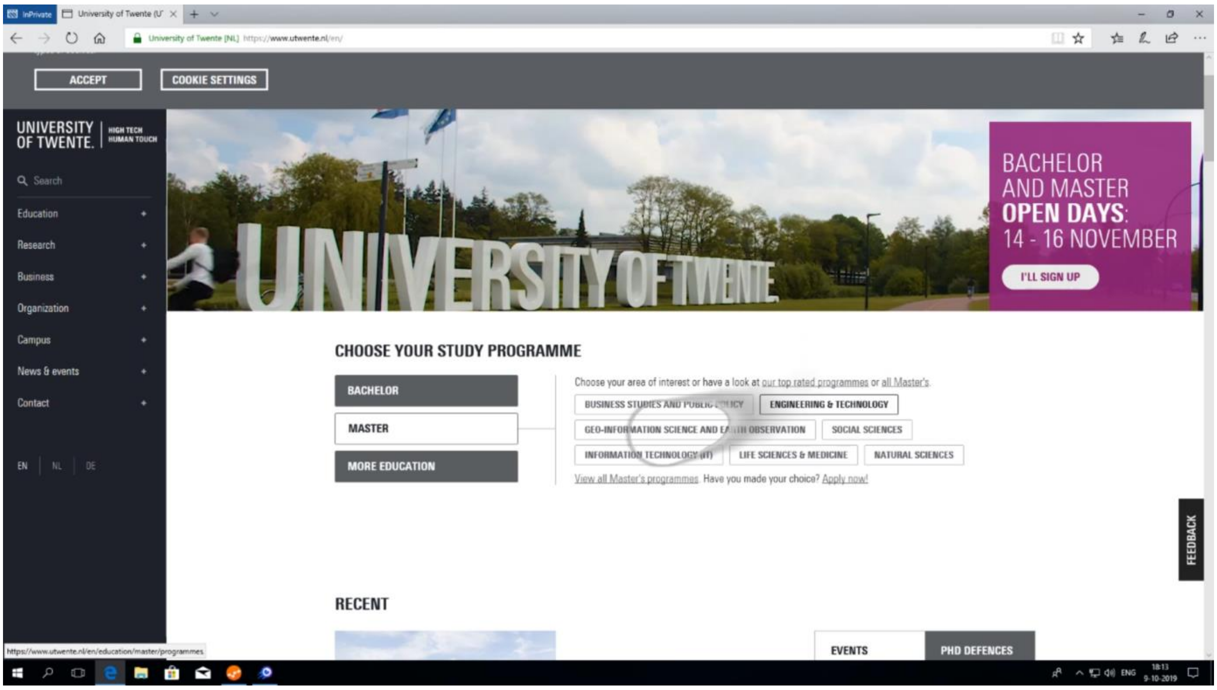

ACCEPT
COOKIE SETTINGS
UNIVERSITY OF TWENTE. HIGH TECH HUMAN TOUCH
Search
Education
Research
Business
Organization
Campus
News & events
Contact
EN | NL | DE
UNIVERSITY OF TWENTE
BACHELOR AND MASTER OPEN DAYS: 14 - 16 NOVEMBER
I'LL SIGN UP
CHOOSE YOUR STUDY PROGRAMME
BACHELOR
MASTER
MORE EDUCATION
Choose your area of interest or have a look at our top rated programmes or all Master's
BUSINESS STUDIES AND PUBLIC POLICY
ENGINEERING & TECHNOLOGY
GEO-INFORMATION SCIENCE AND EARTH OBSERVATION
SOCIAL SCIENCES
INFORMATION TECHNOLOGY (IT)
LIFE SCIENCES & MEDICINE
NATURAL SCIENCES
View all Master's programmes. Have you made your choice? Apply now!
FEEDBACK
RECENT
EVENTS
PHD DEFENCES

Figure 1. The homepage of the website of the University of Twente with vision bubble generated by the eye tracking device Tobii 4C. The vision bubble is grey of colour, and this vision bubble in this still figure is moving from the right to the left of the screen. Adapted from video recordings of the present study.

This study is a within-subject design, because every participant will go through the four different conditions. With every condition, a task is assigned to test this condition. In order to diminish the risk of creating biases, the order of the four different conditions is randomised. This randomisation is further explained in the procedure section.

One drawback from the concurrent think-aloud protocol is that the CTAP can suffer from reactivity from users. A way to diminish the reactivity is by making use of the three levels of verbalisation by Ericsson and Simon (1993). The three levels of verbalisation are three methods that researchers use to communicate with participants during the studies. The difference of the methods is at which cognitive level the communication between researcher and participant takes place. Usually, the communication between researcher and participant is led by the researcher through asking relevant questions to the participant. The first level is based on the short term memory that is verbally encoded. An example of a question that the researcher can ask from the first level is *"Which word are you reading right now?*" The second level is based on the short term memory that is not verbally encoded and a simple cognitive operation. An example of a question that the researcher can ask from the second level is "*What do you see on the screen?*" The third level is based on the long term memory and a complex cognitive operation. An example of a question that the researcher can ask from the third level is "*Which steps where necessary to find this page on the website?*" The idea behind this diminishing of the reactivity is that so long the communication between researcher and participant is either the first or second level, the communication is useful and harmless. In the case that the communication is as according to third level, the communication can be possibly reactive (Ericsson & Simon, 1993). Therefore the researchers from present study have only used first or second level verbalisation in the communication with the participants.

4.1.1 Participants

In total, nineteen people participated in the pilot. The age of the participants ranged from 20 to 29 years (M = 23.11, SD = 2.23). The number of male participants was 11. There were eleven participants with a German nationality, seven participants with a Dutch nationality, and one participant with a Bulgarian nationality. The participants were recruited using a convenience sample.

4.1.2 Apparatuses and Materials

In this study, several different pieces of apparatuses are used. Two eye trackers are used, namely the Tobii Pro X3-120 and the Tobii 4C. The Tobii Pro X3-120 is used for the two classic conditions. This eye tracker has as advantage that it can collect metrics while being used by analysing the gaze of its user. An example of metrics is the time a participant is looking at a certain area of interest. This Pro eye tracker is specifically used in research. The Tobii 4C on the other hand is used for the two cued conditions, and has as advantage that it can relay the gaze of its user in real time by creating a vision bubble on the screen which can be seen in Figure 1. Therefore this eye tracker is often used in gaming and streaming games online and makes users of the vision bubble aware of where their gaze is. This awareness is the cue that stimulates the participants to formulate in more detail the potential usability problems they have with the tested system (Tobii Technology, 2009). The other apparatuses used in this study are a multitude of computers. One computer is used during the experimental phase, for participants to do the tasks and collecting the data generated by the computer. This computer contains the software and corresponding licenses in the BMS-lab of the University of Twente. Another computer is used to analyse the data and write a report around the data. As internet browser to make the tasks doable, Microsoft Edge version 44.18362.267.0 is used.

In order to record the screen while a participant is working on a task, the software Tobii Pro Lab that works with the eye trackers is used. The generated footage is used in the retrospective think-aloud protocol and for the data analysis. Besides a screen recorder, an audio recorder is used for recording the comments made by the participants.

As materials, this study uses five questionnaires. One questionnaire focuses on the demographic information, and can be found in the Appendix C. This questionnaire consists of three open questions and five multiple-choice questions. The other four questionnaires focus on the participants' experience and opinions about the condition the participants have been using, and can be found in Appendix B. There is one questionnaire for every condition, because the two retrospective conditions get two additional questions regarding the playback video, and the two cued conditions get two additional questions regarding the vision bubble. Every questionnaire makes use of a 5-point Likert scale, in which 1 equals 'Strongly disagree' and 5 equals 'Strongly agree'. This will be presented to the participant after each condition. Besides the questions that participant must answer, the participant also have some space to put any comment the participant still wants to give.

4.1.3 Tasks

A task environment is necessary to operate the four conditions. The present study uses a task environment found close by home, namely on the recruitment website for master studies from the University of Twente. The web address of the recruitment website for master studies is https://www.utwente.nl/en/education/master/. This website contains information about the different master studies that are offered at this university. Students that are interested in doing a master at the University of Twente can browse and search for specific information about the master the students are interested in. Therefore this study is interested in participants with an interest in doing a master at the University of Twente, and participants who have chosen to do a master at the University of Twente. The research team has collaborated with the Marketing & Communication department of the University of Twente, to design the tasks that the participants will do. Four tasks were designed by the Marketing & Communication department together with the research team, and have been transformed into four different scenarios. The four scenarios that contain the tasks can be found in an overview in Appendix D, including the assignment to the condition. The results regarding the usability problems that are found on the website of the University of Twente will be shared with the Marketing & Communication department after the study, in order to improve the website.

4.1.4 Procedure

Before a participant starts the study, the order of the four different conditions is randomised to counter the creation of biases. The randomisation is happening in two steps. First, it is randomly decided if the participant starts with the classic conditions or the cued conditions. This means that the researcher have to change the eye tracking device once, in order to diminish the risk of errors. The second step is to decide whether the participant starts with the concurrent think-aloud protocol or the retrospective think-aloud protocol. The randomisations occur with help from the website random.org. This website creates certified true randomness by using atmospheric noise (random.org, n.d.). Therefore the randomisations contain no biases from predictable algorithms. The randomisations are decided by connecting the numbers one and two to the different conditions, and the numbers three and four to the different protocols. With this connection and the ‘Integer Generator’ from random.org, the randomisations can be made. First, the order of the two conditions is settled by putting the minimum on 1 and the maximum on 2 in the ‘Integer Generator’. By hitting the button ‘RANDOMIZE’ the number one or two will be generated randomly. This number corresponds to one of the two conditions, and therefore the corresponding condition will be tested as first and the other as second. In this case, number 1 stands for ‘Classic’ and number 2 stands for ‘Cued’. This will also happen with the decision of the order of protocols, but with this randomisation the minimum is 3 and the maximum is 4. In this case, number 3 stands for ‘CTAP’ and number 4 stands for ‘RTAP’. The randomisation of the order of protocols happens twice, one time for the classic condition and one time for the cued condition.

After the randomisations are done, the lab is set up for the next participant. This means that the right software is selected and the right eye tracker is prepared. After the set-up is correctly done, the participant is invited into the lab and is asked to fill in an informed consent and the first questionnaire. This questionnaire contains questions focusing on the demographic information, and therefore is interested in for example age, sex, and usage of internet. According to the General Data Protection Regulation (GDPR), private information such as demographic information needs to be processed and treated with care (European Commission, 2018). To ensure the safety of participants, the data is made anonymous so that no person can trace back the data to any of the participants. Furthermore, every hardcopy and digital data is stored behind a lock. After finishing this study, all sensitive information will be destroyed. The eye tracker for the first two conditions is calibrated with the gaze of the participant, and then the trial can start. During the trails, the researcher is encouraged to ask questions if necessary, as is done through the question-asking protocol.

An example of a randomised order of conditions is that the participant starts with the classic CTAP, followed by classic RTAP, followed by cued CTAP, and ending with the cued RTAP. As mentioned, the classic conditions works with the Tobii Pro X3-120 and the cued conditions with the Tobii 4C. With the concurrent think-aloud protocol, participants will speak aloud about their actions and thoughts while working on a certain task. This differs from the retrospective think-aloud protocol because in this protocol participants work in silence on the tasks and can comment afterwards on the participants' performance. The difference between the classic and cued condition is that in the cued condition participants can consciously see where the participants are looking at, thus participants become aware of their own gaze. This awareness is not present at the classic conditions. This creates the four conditions of this study, which are all experienced by every participant. After every task with its corresponding condition, the participant is presented with a questionnaire to assess the experience and opinion of the participant about the task. If the questionnaire has been filled in, then the next condition is prepared by the researcher. After the next condition is fully prepared, the participant can work on the task that corresponds with this condition. After the last task and thus the last questionnaire, the participant is debriefed and thanked for his or her participation.

4.1.5 Data Analysis

The first step of analysing the data generated by the participants is to watch the screen footage and listen to the audio recordings of each trial. By experiencing all the recordings again, the usability problems that the tested participants have with the tested website can be assessed. This is done by noting down every incident that the participants had while working on the tasks. These individual incidents are then matched and organised into groups of incidents that are similar in hindrance or outcome. These groups of incidents will then be known as usability problems. This is done for every task.

After all the usability problems that the participants have with the website have been found, the severity, the types, and the detection methods of the usability problems are assessed by the research team. This creates a thorough understanding of the usability and usability problems of the tested website from the University of Twente. A next step would be to start the comparative analysis with the six comparative criteria and Tobii Pro Lab, used to further examine the differences between conditions. However, this is not possible due to the error in the randomisation.

4.2 Results of the Pilot

An error has occurred in the randomisation making the majority of the data from the pilot study non-comparable with each other. Nonetheless, the pilot study has generated data that is deemed to be usable, but than for the second study. Therefore, this usable data will be discussed in the current results section.

The first part of the usable data from the pilot study concerns the fact that the pilot study had a lack of standardised tests, especially concerning the ease of use of the website that had been tested. Participants have both mentioned on occasion the ease and difficulty to use the website during the pilot study. Therefore, a standardised test or questionnaire could help to clear up the ease of use.

The next part of the usable data from the pilot study concerns the mental effort from participants during the pilot study. Participants have mentioned that with certain tasks and certain conditions, the participants experienced an increased effort than in comparison to the other tasks and conditions. This increased effort was mostly straining the mental capacity of the participants. Therefore, a test or instrument that measures the mental effort of participants during an exercise or task could help clear up about the mental effort that participants experienced.

The last part of the usable data from the pilot study concerns the similarity between the tasks 1 and 4 from the pilot study. After the mistake that was earlier discussed was discovered, the research team has done a task analysis to discover the similarities between the tasks in the first study. An overview can be seen in Table 3. From this overview, it can be seen that task 1 and task 4 are similar. That is also noticeable in Appendix E, which is the overview of the four tasks with description and the required steps to complete a task. Namely in Appendix E at task 1 and task 4, the first seven steps to complete the task are equal. Thus task 1 and task 4 share similarities. These shared similarities can have as effect that participants are biased in completing task 4 if task 1 is before task 4 in the order that participants complete tasks, and biased in completing task 1 if task 4 is before task 1 in the order that participants complete tasks. Therefore this bias is that participants already have knowledge about completing task 4 if the participants have done task 1 before task 4, and knowledge about completing task 1 if the participants have done task 4 before task 1.

Table 3

Overview of the Results of the Task Analysis.

Task:	Steps:		Time on tasks:		Usability problems:
ID	Number	Total in seconds (s) for all participants	Average per participants in seconds (s)	Average: participant per step in seconds (s)	Total number
1	9	4954 s	260.74 s	28.97 s	31
2	9	2685 s	141.32 s	15.7 s	16
3	9	7227 s	380.37 s	42.26 s	64
4	9	4522 s	238 s	26.44 s	42

4.3 Discussion and Lessons Learned from the Pilot

From the three parts of usable data of the pilot study, three adjustments can be made for the second study. The first adjustment is adding a standardised questionnaire that tackles the questions regarding the ease of use of the website that has been tested in the pilot study. One such questionnaire is the UMUX-lite. The questionnaire UMUX-lite, which is derived from the UMUX questionnaire, measures the perception from users on the ease of use of the system the users are working with. The UMUX-lite is a standardised questionnaire, with high internal reliability and high correlation with other standardised questionnaires such as the SUS (Sauro, 2017). Therefore this questionnaire will be used in the second study. More information on the UMUX-lite can be found in methods section of the second study.

The second adjustment is adding an instrument that can enlighten the mental effort from the participants. Such an instrument is the Rating Scale Mental Effort. This is a scale between 0 and 150 in which participants can indicate what their mental effort with help from nine anchor points on the scale. More information on the Rating Scale Mental Effort can be found in the methods section of the second study.

The final adjustment is concerning the similarity of task 1 and task 4. To make sure that participants complete task 1 and task 4 with as less bias as is possible, the order of the tasks in the second study is altered into a quasi-random order. That means that either task 1 or task 4 will be the first task in the order of tasks, and the other task is the last task in the order of tasks. For example, a participant can have task 4 as the first task, and task 1 as the last task. This modification in the second study will diminish the bias of the shared similarities as much as possible.

5. The Comparative Study

5.1 Method of the Comparative Study

The initial study contains a mistake that made the collected data unusable. The mistake was that the each of the four tasks was assigned to one of the methods that are in the interest of the current study. Therefore, to compare the four methods was impossible since every comparison and difference is influenced by the fact that the result of every method is based on just one task. Would the assignment of task and method have not existed, then the comparisons and differences of the methods from the initial study would have a legitimate base. Thus, a second study that is similar to the initial pilot study but without the mistakes could suffice to repair the damage that has been done by the first study.

The second study also makes use of a 2X2 mixed design with the same conditions and tasks as the pilot; the conditions can be found in Table 2 and the tasks can be found in Appendix D. Each participant performed all the tasks (within participants) with one of the possible conditions (between subjects). However, each condition is not attached to one of the tasks, as was done in the pilot study.

5.1.1 Participants

Another difference between the first study and the second study are the participants. In the comparative study, participants were recruited that did not participate in the pilot study. A total of 20 new participants were recruited. The age of the participants ranged from 17 to 25 years (M= 21.55, SD =2.29). The number of male participant was 13. There were ten participants with a German nationality, eight participants with a Dutch nationality, one participant with an American nationality, and one participant with a Lithuanian nationality. The participants were recruited using a convenience sample.

5.1.2 Apparatuses and Materials

The apparatuses and materials used in the comparative study are equal to the apparatuses and materials used in the pilot. That means that the comparative study makes use of the eye tracking device Tobii Pro X3-120 and the eye tracking device Tobii 4C. Furthermore, in the comparative study multiple computers are used; one computer is used during the experimental phase with the software and corresponding licenses in the BSM-lab and one computer is used for the analysis of the data and writing the report. The same internet browser is used, namely the browser Microsoft Edge version 44.18362.267.0. The same recording devices for both video and audio are used in the comparative study as in the pilot.

As for the materials used in the pilot, the comparative study makes use of the same materials. That means that the one demographic questionnaire, which can be found in Appendix C, and the four questionnaires regarding participants' experience, which can be found in Appendix B, is again used in the comparative study. However, the four questionnaires regarding the participants' experience are used in a different way. In the comparative study, a participant only uses one condition for all the four tasks instead of all the four conditions on all the four tasks. Therefore, each participant in the comparative study fills in only one of the four questionnaires regarding the participants' experience.

Furthermore, an additional questionnaire and an additional instrument will be used in the comparative study. Both additions are based on what has been found in the pilot study and therefore is added to the comparative study. The explanation of the basis of the two additions can be found in the Results section of the pilot. The first addition is better known as the UMUX-lite questionnaire. This is a shortened version from the UMUX (Usability Metric for User Experience), and both versions measures the perception of the ease of using a system such as a website (Sauro, 2017). The difference between UMUX and UMUX-lite is that the lite version is shorter and only consists of positive worded items. The benefit of having only positive worded items are that it will create a one-dimensional structure, instead of a bi-dimensional structure (Lewis, Utesch, & Maher, 2015). A one-dimensional structure is beneficial due to the fact that it is less ambiguous in comparison to a bi-dimensional structure. Therefore, a higher score on the UMUX-lite questionnaire means that a system is perceived as easier to use. The UMUX-lite consists of two items, namely 'The system's (website) capabilities meet my requirements' and 'The system (website) is easy to use'. For both items, participants can fill in a seven-point Likert scale ranging between 'Strongly disagree' for 1 and 'Strongly agree' for 7. The questionnaire for the UMUX-lite can be found in Appendix F. The second addition is better known as the Rate Scale Mental Effort (RSME). The RSME is a scale from 0 to 150 in which participants can indicate what their cognitive workload was during a task they just did. On the scale there are nine anchor points that guides the participants in deciding what their mental effort was, as can be seen in Appendix G. Participants can write their absolute rating of mental effort on the form itself.

5.1.3 Procedure

The procedure is also different in the second study than in comparison to the first study. Before any participant started with the second study, the randomisation is completed. This means that a table is made, which contains for each participant the usability technique the participant is going to work with and what the order of the four tasks is. This table and the randomisation process are reported in the Appendix H.

Before the participant enters, the setup of the laboratory was adapted to the type of technique that was randomly assigned to each participant. The next step is to invite the participant into the laboratory and give the participant the informed consent and demographic questionnaire. Before the participant can fill it the two given forms, the general idea behind the study will be explained by the researcher. After the first explanation, the participant will fill in the informed consent and demographic questionnaire. Then the details of the study will be explained, such as what think-aloud protocol is and what kind of different think-aloud protocols exist. At the end of the explanation, the participant will be asked if the participant have any questions. In the case there are still some questions, the researcher will answer the questions where possible. If there are no further questions left, the experiment will start.

The first thing that is done is explained to the participant which technique will be used by the participant. The next step is to calibrate the eye tracking device onto the eyes of the participant. This calibration process is different per eye tracking device. For the Tobii Pro X3-120, the calibration process requires participants to look and follow a dot on the screen which will pause at different spots. After this is done, Tobii Pro Lab will show a results screen from the calibration. If the researcher is satisfied, the experiment can start. For the eye tracker Tobii 4C, the calibration process requires participants to blow up a series dots on the screen if the participant looks long enough at the dot. This is reiterated until the calibration program is satisfied. Thereafter, participants can start the experiment. The participant will work on the first task that is in the order of tasks for this participant. After the first task is done, the participant receives the questionnaire UMUX-lite and the instrument RSME. The participant will fill in the questionnaire and instrument, and then continue with the second task. Again, after the second task the participant will fill the questionnaire UMUX-lite and instrument RSME and then continues with the third task. The questionnaire UMUX-lite and instrument RSME will also be received and filled in by the participant after the third and fourth task. The experiment is ended with a questionnaire about the participant’s experience. After this questionnaire the participant is thanked and debriefed.

5.1.4 Data Analysis

The analysis of the second study is similar to the first study. That means that the same six comparative criteria are used in the second study, as is the analysis of the data from the software Tobii Pro Lab. Besides these six criteria and the data from Tobii Pro Lab, the analysis will consists of the analysis of the questionnaire UMUX-lite and the instrument RSME. From the questionnaire UMUX-lite, a score can be calculated. The equation of the score is equal to the equation of the SUS score, which is the following: SUS Score = 0.65 * ((UMUX-Lite Item1 + UMUX-Lite Item2 – 2) * (100/12)) + 22.9. In this equation, UMUX-lite item 1 is 'The system's (website) capabilities meet my requirements', and UMUX-lite item 2 is 'The system (website) is easy to use'. The SUS (System Usability Scale) score is another way to measure the perceived ease of use of systems. The instrument RSME is on a scale between 0 and 150. The mean for every technique will be compared to each other to see if one technique took more mental effort by the participants than the other.

This study relies on Bayesian regression statistics to analyse all the data that has been gathered and processed. One advantage of this kind of statistics is that it can utilise different families of statistics which are applicable on different types of data (Buerkner, n.d.; Schmettow, 2020). The families used in this study are exGaussian, Cratio, Binomial, Poisson, and Acat. The family exGaussian is used for data that does not assume that the data can be zero. An example of this is the time that a participant has spent on a task. The family Cratio is used for rating scales, such as Likert scales. The family Binomial is used for count data that have a clear upper limit. An effect in the outcome of using the family Binomial is that the outcome is represented between -1 and 1. Oppositely, the family Poisson is used for count data that have no apparent upper limit. The last family is the family Acat, which is used for data that has adjacent categories. The data of the detection methods of usability problems uses in this study the family Acat.

5.2 Results of the Comparative Study

Before the analyses start, all the raw data is processed into data that is workable with. For example, a list of all the usability problems is made by experiencing the trials again by listening and watching the audio and video recordings, as is explained in the methods section.

One aspect of the Bayesian regression statistics is that the output of the statistics is a difference-table. This difference-table works with a reference condition, in which the first condition in the table is the base level on which the other conditions are compared to. Therefore, one criterion may have multiple difference-tables, since there are a total of four different conditions, and thus four different possible reference conditions. A difference between the reference condition and the other condition is significant with 95% if the columns ‘lower’ and ‘upper’ in the difference-table is either both positive or negative. Namely, the column ‘lower’ gives the score for the 2.5% certainty quantile and the column ‘upper’ gives the score for the 97.5% certainty quantile (Schmettow, 2020). Therefore, the focus is to achieve the significance of the data while using the columns ‘lower’ and ‘upper’.

5.2.1 Performance

The performance of the usability tests will be assessed by the time that participants spent on the task and by the fact whether participants have completed the task in the correct way. For the time on task, the two differences that are significant are between the conditions classic CTAP and cued CTAP, and the conditions classic RTAP and cued CTAP, which can be seen in Table 4. The condition that participants spent the least amount of time on tasks is the condition classic RTAP, after that the condition classic CTAP, and as last is the condition cued CTAP. The differences between the condition cued RTAP and the other three conditions are not significant on time on task. The differences between the four conditions on the completion of tasks are not significant.

Table 4

the Difference-Table of the Time on Task with the Condition Cued CTAP as Reference Condition

fixef	center	lower	upper
Cued CTAP	318.84177	238.1356	408.16131
Cued RTAP	-81.63073	-205.3023	11.52348
Classic CTAP	-97.89717	-206.9346	-11.62335
Classic RTAP	-151.02440	-270.1579	-57.39890

5.2.2 Rating Scale Mental Effort

From the instrument regarding the Rating Scale Mental Effort (RSME) arise several significant differences. Both conditions classic and cued RTAP are significantly less demanding than the concurrent think-aloud protocols, with the condition classic RTAP having the lowest demanding in terms of mental effort. The condition cued CTAP is the most demanding. This can be seen in Table 5 for the classic RTAP, in Table 6 for the cued RTAP, and in Figure 2 for all conditions. Figure 2 gives a graphical overview of the scores on the RSME per condition, with the frequency of the scores on the y-axis. For the analysis of the RSME is the family Binomial used, therefore the outcomes in Table 5 and Table 6 are represented between -1 and 1.

Table 5

the Difference-Table of the Rating Scale Mental Effort Score with the Condition Classic RTAP as Reference Condition

fixef	center	lower	upper
Classic RTAP	-0.5973714	-0.6902247	-0.5087761
Cued CTAP	0.8053561	0.6785671	0.9370457
Cued RTAP	0.4696016	0.3441685	0.5957699
Classic CTAP	0.7645841	0.6348455	0.8907566

Table 6

the Difference-table of the Rating Scale Mental Effort Score with the Condition Cued RTAP as Reference Condition

fixef	center	lower	upper
Cued RTAP	-0.1256209	-0.2136758	-0.0496820
Classic CTAP	0.2948180	0.1776485	0.4144351
Classic RTAP	-0.4713597	-0.5911654	-0.3493602
Cued CTAP	0.3372073	0.2188126	0.4540186

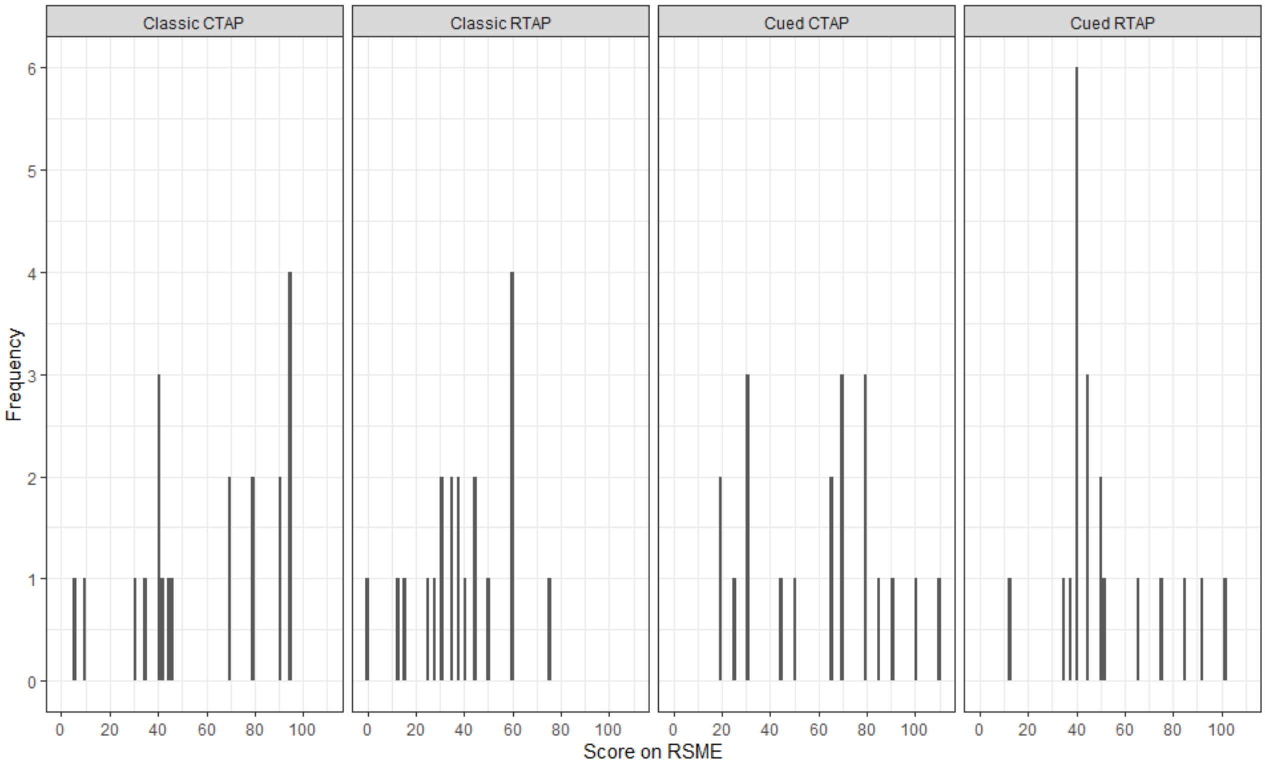


Figure 2. The graphical overview of the scores on the RSME per condition, with the frequency of the scores on the y-axis. Some lines may appear to be thicker in width than others, due to scores that are close to each other.

5.2.3 UMUX-lite

From the questionnaire regarding the UMUX-lite arise several significant differences. Both conditions cued CTAP and cued RTAP score significantly lower than the classic conditions, with the condition cued CTAP having the lowest score on the UMUX-lite. The condition classic RTAP has the highest score on the UMUX-lite. This can be seen in Table 7 for the cued CTAP, in Table 8 for the cued RTAP, and in Figure 3 for all conditions. Figure 3 gives a graphical overview of the scores on the UMUX-lite per condition, with the frequency of the scores on the y-axis. For the analysis of the UMUX-lite is the family Binomial used, therefore the outcomes in Table 7 and Table 8 are represented between -1 and 1.

Table 7

the Difference-Table of the UMUX-lite score with the Condition Cued CTAP as Reference Condition

fixef	center	lower	upper
Cued CTAP	0.2885579	0.1990643	0.3826962
Cued RTAP	0.5869258	0.4506907	0.7214925
Classic CTAP	0.7697270	0.6298229	0.9085608
Classic RTAP	0.7876588	0.6400739	0.9279535

Table 8

the Difference-Table of the UMUX-lite score with the Condition Cued RTAP as Reference Condition

fixef	center	lower	upper
Cued RTAP	0.8761239	0.7746467	0.9789273
Classic CTAP	0.1822098	0.0333373	0.3345851
Classic RTAP	0.1997837	0.0492617	0.3482907
Cued CTAP	-0.5877446	-0.7224421	-0.4497964

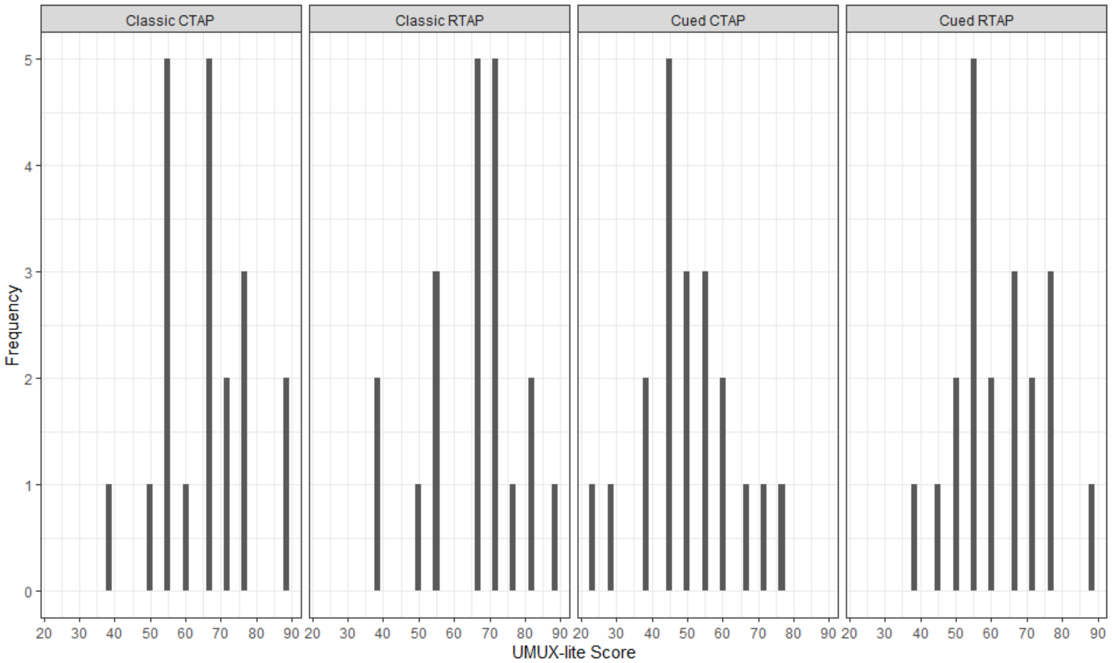


Figure 2. The graphical overview of the scores on the UMUX-lite per condition, with the frequency of the scores on the y-axis.

5.2.4 Participants' Experience

As is explained in the methods section, the participants' experience is assessed through a questionnaire consisting of 15 items. Out of the 15 items from the questionnaires, there are six items with significant differences. The difference-tables from the six items with significant differences can be found in Appendix I.

- Item 'I found it unnatural to verbalise my thoughts': in this item, participants found it significantly more unnatural to verbalise their thoughts with the condition classic CTAP than with the condition cued RTAP.
- Item 'I found it unpleasant to verbalise my thoughts': in this item, participants found it significantly less unpleasant to verbalise their thoughts with the condition cued CTAP than with the condition classic CTAP.
- Item 'I found it tiring to verbalise my thoughts': in this item, participants found the conditions classic CTAP and classic RTAP significantly more tiring that the cued conditions, with the condition classic CTAP to be the most tiring. For the cued conditions, the condition cued RTAP was more tiring than the condition cued CTAP.

- Item ‘I found it time-consuming to verbalise my thoughts’: in this item, participants found the condition classic CTAP to be the significantly most time-consuming, after that the condition classic RTAP, and as last is the condition cued CTAP. The differences between the condition cued RTAP and the other three conditions are not significant on this item.
- Item ‘the links in the texts help me to easily find more information on specific subjects’: in this item, participants experienced that the links in the text with the condition cued RTAP helps the participants significantly more easily to find more information on specific subjects than with the condition cued CTAP.
- Item ‘the content on the master programme page persuades me to read more about the programme’: in this item, participants experienced that the content on the master programme page persuades them significantly less to read more about the programme with the condition classic RTAP, and after that the condition classic CTAP, and as last the condition cued CTAP. The differences between the condition cued RTAP and the other three conditions are not significant on this item.

5.2.5 Usability Problems

There are no significant differences identified in terms of the amount of usability problems, the severity levels of the usability problems, the detection method of the usability problems, and the types of usability problems.

5.2.6 Data Gathered Through Tobii Pro Lab

By using the Tobii Pro Lab, the count of fixations can be assessed. By using the condition classic RTAP, the count of fixations from the participants will be significantly less than in comparison to the condition classic CTAP. This can be seen in Table 9. The differences between the two conditions on the count of fixations in verbalised usability problems are not significant.

Table 9

the Difference-Table of the Amount of Fixations with the Condition Classic CTAP as Reference Condition

fixef	center	lower	upper
Classic CTAP	6.1932669	6.1734067	6.2122166
Classic RTAP	-0.0558065	-0.0839517	-0.0266277

5.2.7 Summary of the Results

In Table 10, an overview of the results of the comparative study with ranking under the light of the comparative six criteria and additional questionnaire and instrument is shown. In this table, the four conditions are ranked per criterion from the condition that fits more with the criterion to the condition that fits less with the criterion. The condition that fits the most with the criterion is given the ranking-score of 4, while the condition that fits the least with the criterion is given the ranking-score of 1. The ranking-scores of 2 and 3 are for the conditions that are respectively the second least fitting condition and second most fitting condition. The criteria concerning the amount of usability problems, the severity level of the usability problems, the types of usability problems, and the methods how the usability problems have been detected are not significantly different on the four conditions. Therefore, these criteria are not found in Table 10. Furthermore, the criterion concerning the count of fixations assessed through Tobii Pro Lab is not found in Table 10. The reason for this is that this criterion is only measured in the two classic conditions.

Table 10

an Overview of the Results of the Comparative Study with Ranking-Scores under the Light of the Comparative Six Criteria and Additional Questionnaire and Instrument that were Significantly Different

Criteria	Classic		Cued	
	CTAP	RTAP	CTAP	RTAP
Performance: Time on Task.	3	4	2	1
Participants’ experience.	1	2	4	3
Rating Scale Mental Effort	2	4	1	3
UMUX-lite	3	4	1	2
Total	9	14	8	9

6. Discussion of the Comparative Study

The results from the present study suggest that the condition classic RTAP fits the criteria better than the three other conditions. Furthermore, the results suggest that using the two cued conditions fit the criterion regarding the participants' experience better than the two classic conditions. However, the two cued conditions fit the criteria regarding the questionnaire UMUX-lite and the time spent on a task worse than the two classic conditions.

These suggestions and the other results can differ or be in line from what is known from previous research. To start with the criterion regarding performance, participants working with the condition cued CTAP take significantly more time with finishing a task than in comparison to the conditions classic CTAP and classic RTAP. This result is not in line with previous research. An article from Eger, Ball, Stevens, and Dodd (2007) gives insight into the differences and similarities between using cues and not using cues in think-aloud protocols. One result of their study is that the time taken to complete the primary task has no differences at all. One explanation would be that the vision bubble that cues the participants in the condition cued CTAP would increase the cognitive workload of participants, on top of the increased cognitive workload from the Concurrent Think-Aloud Protocol. This can result in reactivity and an increased amount of time spent on a task by the participants, which has occurred in the present study. This is in line with the result of the Rating Scale Mental Effort from the present study. This result is that participants working with both conditions classic CTAP and cued CTAP admitted to experiencing more cognitive strain than in comparison to the two conditions classic RTAP and cued RTAP. According to the participants, the condition that required the least amount of mental effort was the condition classic RTAP, and after that the condition cued RTAP. Thus, according to the present study, the CTAP conditions and the cued conditions require the most mental effort, and therefore the most time per task.

With the criterion regarding the UMUX-lite, participants working with the conditions classic CTAP and classic RTAP have a higher score on the UMUX-lite score than in comparison to the conditions cued CTAP and cued RTAP. With the two cued conditions, the condition cued CTAP has a lower score than the condition cued RTAP. What this means is that participants working with the conditions classic CTAP and classic RTAP consider a system, such as a website, to be more easy to use and the capabilities of this system meet more of the requirements set by the participants. One explanation of this result is similar to what is discussed in the previous two criteria. The vision bubble of the two cued conditions can have as effect that the cognitive workload is increased. This increased workload can then has as effect that participants consider a system’s capabilities to meet less requirements set by the participants and a system less easy to use. Unfortunately, no previous research has been done that combined the UMUX-lite questionnaire with Think-Aloud Protocols, thus no comparison with the present study and other research can be made.

In terms of criterion regarding participants’ experience, participants working with the condition classic CTAP experienced that verbalising their thoughts was more unnatural than in comparison to the condition cued RTAP. This same result of the present study is not found in previous research from Alhadreti and Mayhew (2018), Van den Haak, De Jong, and Schellens (2003), and Eger, Ball, Stevens, and Dodd (2007). One explanation about the difference between the current study and literature is that the two conditions that are significantly different in the current study have both the think-aloud protocols and the eye tracking devices different. For the think-aloud protocol, the condition classic CTAP is considered to be more unnatural because participants feel more unnatural when doing both thinking aloud and working on a task, while participants that solely doing thinking aloud feel more natural. For the difference in cueing due to the different eye tracking devices, the condition classic CTAP is considered to be more unnatural because the vision bubble helps in doing the thinking aloud, therefore seeming more natural. The combination of the differences in think-aloud protocols and eye tracking devices make that the conditions classic CTAP and cued RTAP are significantly different.

Furthermore with the criterion regarding participants’ experience, participants working with the condition cued CTAP experienced that verbalising their thoughts was less unpleasant than the participants working with the condition classic CTAP. This result is also found in another study. In a study from Eger, Ball, Stevens, and Dodd (2007), the researchers examined the differences between cueing and not cueing while using think-aloud protocols. Participants from that study found that not cueing is significantly more unpleasant than cueing the participants. One explanation of this similarity in results is that cueing participants can help the participants in completing the tasks, thus cueing feels less unpleasant to the participants.

Also with the criterion regarding participants’ experience, participants found working with the two classic conditions to be more tiring than the two cued conditions. This result is not found in the study from Eger, Ball, Stevens, and Dodd (2007). In that study, the difference between cueing and not cueing participants does not have a significant effect on whether participants found it tiring to verbalise their thoughts. One explanation for this significant difference of the current study is that with the two classic conditions, participants do not receive any guidance to do usability testing besides their own thinking mind. With the two cued conditions, the advantage for participants is that they are helped with cues that guide the participants to do usability testing. Therefore participants working with the conditions classic CTAP and classic RTAP must do all the work on their own, and thus making this work more tiring.

Besides with the criterion regarding participants' experience, participants working with the condition classic CTAP found it more time-consuming to verbalise their thoughts than with the other three conditions. Looking at previous research, the studies from Van den Haak, De Jong, and Schellens (2003), and Eger, Ball, Stevens, and Dodd (2007) have found that the item about time-consuming is not significant. The study from Alhadreti and Mayhew (2018) however, has found that the retrospective condition is significantly more time-consuming than the concurrent condition. One explanation for the fact that the current study considers the condition classic CTAP to be more time-consuming is that participants underestimate the CTAP. As mentioned in the procedure of the methods section, every think-aloud protocol is explained to the participant. This can give a certain expectation from the participants about the different protocols, where the concurrent protocol is considered to be shorter than the retrospective protocol. The pitfall for participants about this expectation, which can have as a result that the participants consider the condition more time-consuming, is that thinking aloud while working on a task is not as time efficient as only working on a task. Therefore the participants underestimate the condition classic CTAP and consider it in the end to be more time-consuming.

Additionally with the criterion regarding participants' experience, participants working with the condition cued RTAP considered links in text to help more easily find additional information on specific subjects than the condition cued CTAP. One explanation could be that participants in the condition cued RTAP have a lower strained cognitive workload due to not having to think aloud, and therefore can decide whether a link in the text is worth following through than participants working with the CTAP conditions. Besides that, the vision bubble can help participants to be more aware of useful links in the texts. Unfortunately, no other study has been found that combined the item on links in texts from the questionnaire about the participants' experience with Think-Aloud Protocols, thus no comparison with the current study and any other study can be made.

What’s more with the criterion regarding participants’ experience, participants working with condition classic RTAP are less persuades by the content on the master programme page to read more about the programme than with the two conditions classic CTAP and cued CTAP. One explanation could be that participants working with the condition classic RTAP have a lower strained cognitive workload, therefore participants have the mental capacity to think about the fact whether the content on the master programme page is persuasive or not. While with the two conditions classic CTAP and cued CTAP, participants cannot as easily be persuaded by the content on the master programme page to read more because an increased cognitive workload. Unfortunately, no other study has been found that combined the item on persuasiveness of the content of the master programme page from the questionnaire about the participants’ experience with Think-Aloud Protocols, thus no comparison with the current study and any other study can be made.

In terms of count of fixations, participants working with the condition classic RTAP have less fixations on screen than the condition classic CTAP. One explanation could be that participants working with the classic CTAP are more concentrated on the screen itself due to a higher cognitive workload, because participants are required to work on a task and think aloud at the same time. Furthermore participants in the condition classic RTAP can think silently and look away from the screen, making the count of fixations less than the condition classic CTAP. Unfortunately, no other study has been found that combined count of fixations gathered through Tobii Pro Lab with Think-Aloud Protocols, thus no comparison with the current study and any other study can be made.

6.1 Limitations

Several limitations can have an influence on the interpretation and application of the current study. One limitation concerns the calibration of the two eye tracking devices. With the eye tracker Tobii X3-120, after the calibration is done a results page is shown to assess whether the calibration was successful enough to continue. If not, another calibration is possible. With the other eye tracker, the Tobii 4C, the calibration is finished when the calibration program decides that it is successful. However, a researcher cannot express his opinion on whether the calibration can be called a success, and cannot verify the details on the calibration. Therefore the accuracy of the Tobii 4C eye tracker cannot be guaranteed to match the accuracy of the Tobii X3-120 eye tracker. And a mismatch in accuracy can result in faulty data from the Tobii 4C.

Another limitation concerns the recruitment of participants that are familiar with the website that has been tested during the current study, namely the website of the University of Twente and in specific the master recruitment pages. This limitation has two consequences as effect. The first consequence is that participants can complete the tasks with more ease than participants with less or no knowledge about the website. This consequence is strengthened by the fact that participants from the current study that are in their master studies can already have a larger knowledge about the website of the University of Twente than in comparison to participants that are in their first year of the Bachelor. A research should have recruited participants with differences that are in the interest of the research itself; therefore any additional difference between participants can be seen as unnecessary bias. Unfortunately, the current study has limited resources and therefore was required to recruit participants with difference in prior knowledge of the website.

The other consequence of the recruitment of participants familiar with the website that has been researched in the current study is the attitude about the website from participants. As is mentioned in the method section of the pilot study, the researchers from the current study have collaborated with the Marketing & Communication department from the University of Twente. One of the reasons for the collaboration is an image problem of the website of the University of Twente that appears within the students of the university. Participants have mentioned on the participants' experience questionnaire that they might be biased due to the fact that the participants consider the University of Twente's website to be not good working. Therefore participants might be biased in working with the website and thus may have a less positive experience with the usability testing condition the participants have been working with and the whole study.

6.2 Recommendations for Future Studies

Results from the current study showed that the think-aloud protocols, cueing with a vision bubble, and mental effort have a relationship with each other. This relationship can be a focus of further research, to better understand the difference between concurrent and retrospective think-aloud protocol with regards to mental effort. It also allows to further understanding of the role of cueing within think-aloud protocols.

Another recommendation for future studies is to investigate the differences between the current study and other studies from the literature. These differences are found in the criteria regarding the performance from the participants, and the three items from the questionnaire concerning the experiences from the participants. These three items are about how participants experienced the think-aloud protocols in terms of finding verbalising thoughts unnatural, tiring, and time-consuming.

Another recommendation for future studies is investigating and exploring other combinations of methods and techniques that were introduced in the introduction of the present study. This could result in a better understanding of the different methods and techniques that can be used in usability tests.

7. Conclusion

The purpose of the present exploratory study was to examine the differences and similarities between four conditions that are used in usability testing. Although the four conditions do excel at different aspects, they are also inferior on other aspects. However, the condition classic RTAP is considered to be the condition that fits the criteria better than the three other conditions. Furthermore, the cued conditions fit the criterion regarding the participants' experience better than the two classic conditions. The same two cued conditions fit the criteria regarding the questionnaire UMUX-lite and the time spent on a task worse than the two classic conditions

There is still a lot unknown about usability testing and its methods and techniques. Some studies in the literature agree with the results from the present study, but other studies disagree with the results of the present study. And some results are new to the literature, such as the results regarding the Rating Scale Mental Effort and the UMUX-lite. The present study can be seen as an in-depth exploration into the insights of using think-aloud protocols and eye tracking in usability testing, and encourages investigating think-aloud protocols and eye tracking further with the new knowledge from the present study.

8. References

Alhadreti, O., & Mayhew, P. (2018). Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. *Proceedings of the Conference on Human Factors in Computing Systems, 44,* 1-12. doi: 10.1145/3173574.3173618

Babich, N. (2019). Top 7 Usability Testing Methods. Retrieved from https://xd.adobe.com/ideas/process/user-testing/top-7-usability-testing-methods/

BMS Lab. (n.d.). Working in the Flexperiment rooms. Retrieved from https://bmslab.utwente.nl/knowledgebase/flexperiment-rooms/

Buerkner, P. C. (n.d.) Brmsfamily: Special Family Functions For Brms Models. Retrieved from https://www.rdocumentation.org/packages/brms/versions/2.12.0/topics/brmsfamily

Cambridge Dictionary. (n.d.). [Dictionary entry for the word Technique]. Retrieved from https://dictionary.cambridge.org/us/dictionary/english/technique

Casaló, L. V., Flavián, C., & Guinalíu, M. (2007). The role of security, privacy, usability and reputation in the development of online banking. *Online Information Review, 31*(5), 583-603. doi: 10.1108/14684520710832315

Dumas, J. S., & Redish, J. C. (1999). *A Practical Guide to Usability Testing: Revised Edition.* Exeter, United Kingdom: Intellect Books.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay. *Proceedings of HCI, 21,* 129-137. Retrieved from https://www.scienceopen.com/

Elbabour, F., Alhadreti, O, & Mayhew, P. (2017). Eye Tracking in Retrospective Think Aloud Usability Testing: Is There Added Value? *Journal of Usability Studies, 12*(3), 95-110.

Elling, S., Lentz, L., & De Jong, M. (2011). Retrospective Think-Aloud Method: Using Eye Movements as an Extra Cue for Participants' Verbalizations. *Proceedings of HCI, 25,* 1161-1170. Retrieved from https://dl.acm.org/

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT press.

European Commission. (2018). 2018 reform of EU data protection rules. Retrieved from: ec.europa.eu

Hassan, H., M., & Galal-Edeen, G. H. (2017). From Usability to User Experience. *ICIIBMS, 2018,* 216-222. doi: 10.1109/ICIIBMS.2017.8279761

Hotjar. (2019). The best usability testing methods for your projects. Retrieved from https://www.hotjar.com/usability-testing/methods/

ISO. (2018). ISO 9241-11:2018(en): Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE and AltUsability. *International Journal of Human-Computer Interaction, 31*(8), 496-505. doi: 10.1080/10447318.2015.1064654

Palmer, J. W. (2002). Web Site Usability, Design, and Performance Metrics. *Information Systems Research, 13*(2), 151-167. doi: 10.1287/isre.13.2.151.88

Poole, A., & Ball, L. (2005) Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. *Encyclopedia of Human Computer Interaction* (pp. 211-219). doi: 10.4018/978-1-59140-562-7.ch034

Random.org (n.d.). What’s this fuss about true randomness? Retrieved from https://www.random.org/

Roy, M. C., Dewit, O., & Aubert, B. A. (2001). The impact of interface usability on trust in Web retails. *Internet Research: Electronic Networking Applications and Policy, 11*(5), 388-398. doi: 10.1108/10662240110410165

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*(6), 759-769. doi: 10.3758/BF03202637

Sauro, J. (2017). Measuring usability: from the SUS to the UMUX-lite. Retrieved from https://measuringu.com/umux-lite/

Schmettow, M. (2020). New statistics for the design researchers. Retrieved from https://schmettow.github.io/New_Stats/index.html

Tobii Dynavox. (2020). I-15+. Retrieved from https://www.tobiidynavox.com/devices/eye-gaze-devices/I-15/

Tobii Group. (2020). This is Eye Tracking. Retrieved from https://www.tobii.com/group/about/this-is-eye-tracking/

Tobii Pro Lab. (2020). Tobii Pro Lab User's Manual. Retrieved from https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/Tobii-Pro-Lab-User-Manual/?v=1.142

Tobii Technology. (2009). *Retrospective Think Aloud and Eye Tracking: Comparing the value of different cues when using the retrospective think aloud method in web usability testing* [White paper].

Usability Home (n.d.) Usability Evaluation. Retrieved from http://www.usabilityhome.com/

Van den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2003). Retrospective vs. Concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-351. doi: 10.1080/0044929031000

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability Engineering: Our Experience and Evolution. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 791 817). doi: 10.1016/B978-0-444-70536-5.50041-5

9. Appendices

9.1 Appendix A: Criteria to Compare Setups on the Differences and Similarities between Conditions with Sources.

Criterion	Sources
Performance	- Alhadreti, Mayhew 2018; - Van den Haak, De Jong, Schellens 2003; - Eger, Ball, Stevens, Dodd 2007.
Amount of usability problems	- Alhadreti, Mayhew 2018; - Tobii Technology 2009; - Van den Haak, De Jong, Schellens 2003; - Eger, Ball, Stevens, Dodd 2007; - Elbabour, Alhadreti, Mayhew 2017; - Elling, Lentz, de Jong 2011.
Severity level	- Alhadreti, Mayhew 2018; - Elbabour, Alhadreti, Mayhew 2017; - Dumas, Redish 1999.
Usability problems types	- Alhadreti, Mayhew 2018; - Tobii Technology 2009; - Van den Haak, De Jong, Schellens 2003; - Eger, Ball, Stevens, Dodd 2007; - Elbabour, Alhadreti, Mayhew 2017; - Elling, Lentz, de Jong 2011.
Detection method	- Elbabour, Alhadreti, Mayhew 2017; - Elling, Lentz, de Jong 2011.
Participants’ experience	- Alhadreti, Mayhew 2018; - Van den Haak, De Jong, Schellens 2003; - Eger, Ball, Stevens, Dodd 2007; - Elbabour, Alhadreti, Mayhew 2017; - Elling, Lentz, de Jong 2011.

9.2 Appendix B: Questionnaire Participants’ Experience

9.2.1 Appendix B1: Classic CTAP

	Strongly agree				**Strongly disagree**
1. I found it difficult to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
2. I found it unnatural to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
3. I found it unpleasant to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
4. I found it tiring to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
5. I found it time-consuming to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
6. The presence of the researcher was unnatural.	**1**	**2**	**3**	**4**	**5**
7. The presence of the researcher was disturbing.	**1**	**2**	**3**	**4**	**5**
8. The presence of the researcher was unpleasant.	**1**	**2**	**3**	**4**	**5**
9. The links in the texts help me to easily find more information on specific subjects.	**1**	**2**	**3**	**4**	**5**
10. The content on the Master Programme page persuades me to read more about the programme.	**1**	**2**	**3**	**4**	**5**
11. I like the tone of voice that is used in the Master Programme site.	**1**	**2**	**3**	**4**	**5**

Additional comments:

9.2.2 Appendix B2: Classic RTAP

	Strongly agree				**Strongly disagree**
1. I found it difficult to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
2. I found it unnatural to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
3. I found it unpleasant to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
4. I found it tiring to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
5. I found it time-consuming to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
6. Seeing the playback video distracted me in remembering what I thought.	**1**	**2**	**3**	**4**	**5**
7. I disliked seeing the playback video while verbalising my thoughts.	**1**	**2**	**3**	**4**	**5**
8. The presence of the researcher was unnatural.	**1**	**2**	**3**	**4**	**5**
9. The presence of the researcher was disturbing.	**1**	**2**	**3**	**4**	**5**
10. The presence of the researcher was unpleasant.	**1**	**2**	**3**	**4**	**5**
11. The links in the texts help me to easily find more information on specific subjects.	**1**	**2**	**3**	**4**	**5**
12. The content on the Master Programme page persuades me to read more about the programme.	**1**	**2**	**3**	**4**	**5**
13. I like the tone of voice that is used in the Master Programme site.	**1**	**2**	**3**	**4**	**5**

Additional comments:

9.2.3 Appendix B3: Cued CTAP

	Strongly agree				**Strongly disagree**
1. I found it difficult to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
2. I found it unnatural to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
3. I found it unpleasant to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
4. I found it tiring to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
5. I found it time-consuming to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
6. Seeing the gaze-bubble distracted me in remembering what I thought.	**1**	**2**	**3**	**4**	**5**
7. I disliked seeing the gaze-bubble while verbalising my thoughts.	**1**	**2**	**3**	**4**	**5**
8. The presence of the researcher was unnatural.	**1**	**2**	**3**	**4**	**5**
9. The presence of the researcher was disturbing.	**1**	**2**	**3**	**4**	**5**
10. The presence of the researcher was unpleasant.	**1**	**2**	**3**	**4**	**5**
11. The links in the texts help me to easily find more information on specific subjects.	**1**	**2**	**3**	**4**	**5**
12. The content on the Master Programme page persuades me to read more about the programme.	**1**	**2**	**3**	**4**	**5**
13. I like the tone of voice that is used in the Master Programme site.	**1**	**2**	**3**	**4**	**5**

Additional comments:

9.2.4 Appendix B4: Cued RTAP

	Strongly agree				**Strongly disagree**
1. I found it difficult to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
2. I found it unnatural to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
3. I found it unpleasant to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
4. I found it tiring to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
5. I found it time-consuming to verbalise my thoughts.	**1**	**2**	**3**	**4**	**5**
6. Seeing the gaze-bubble distracted me in remembering what I thought.	**1**	**2**	**3**	**4**	**5**
7. I disliked seeing the gaze-bubble while verbalising my thoughts.	**1**	**2**	**3**	**4**	**5**
8. Seeing the playback video distracted me in remembering what I thought.	**1**	**2**	**3**	**4**	**5**
9. I disliked seeing the playback video while verbalising my thoughts.	**1**	**2**	**3**	**4**	**5**
10. The presence of the researcher was unnatural.	**1**	**2**	**3**	**4**	**5**
11. The presence of the researcher was disturbing.	**1**	**2**	**3**	**4**	**5**
12. The presence of the researcher was unpleasant.	**1**	**2**	**3**	**4**	**5**
13. The links in the texts help me to easily find more information on specific subjects.	**1**	**2**	**3**	**4**	**5**
14. The content on the Master Programme page persuades me to read more about the programme.	**1**	**2**	**3**	**4**	**5**
15. I like the tone of voice that is used in the Master Programme site.	**1**	**2**	**3**	**4**	**5**

Additional comments:

9.3 Appendix C: Demographic Questionnaire

Gender:

○ Male

○ Female

○ Other

○ I don’t want to specify.

Date of birth: ......-……-…..…….

Nationality: ....

Current level of education:

○ Bachelor year 1

○ Bachelor year 2

○ Bachelor year 3

○ Master

Current study program: ....

How often do you use an internet browser (e.g. Chrome, Firefox, Safari, etc.)?

○ Several hours per day

○ Several hours per week

○ Several hours per month

○ Several hours per year

○ Never

What is your experience with Think Aloud Protocols on a scale of 1 to 5, where 1 is ‘I have zero experience’ and 5 is ‘I consider myself an expert’?

○ 1

○ 2

○ 3

○ 4

○ 5

What is your experience with eye tracking on a scale of 1 to 5, where 1 is 'I have zero experience' and 5 is 'I consider myself an expert'?

○ 1

○ 2

○ 3

○ 4

○ 5

9.4 Appendix D: Overview of Tasks and Corresponding Conditions

Task Number	Task	Condition
1	Peter, an old friend of yours, would like to follow the master programme Health Sciences from the University of Twente. He cannot momentarily access to the Internet and he asked you to help him by finding out if in the master programme Health Sciences; there is a course that would prepare him to deal with data. Can you help him? Please start your research from the homepage of the university.	Classic CTAP
2	Francis, a friend of yours, recently finished his Bachelor, and he is thinking about a master at the University of Twente. Since you already study at the University of Twente, he has asked you for help. Your friend was wondering if there is a specialization on the maintenance and the operations of mechanical objects that will enable him to graduate as an engineer. Can you help him? Please start your research from the homepage of the university.	Classic RTAP
3	Paula, a friend of yours, has finished her Bachelor Psychology at the University of Twente some time ago. She is interested in doing a master programme in Systems and Control at the University of Twente. However, she is not sure if she can enter this master programme with a Bachelor in Psychology. Can you help her? Please start your research from the homepage of the university.	Cued CTAP
4	Ni, a Chinese friend of yours, is looking to subscribe to the master programme Health Sciences at Twente. He would like to attend the course in September 2020. However, he is not sure how much tuition fees he is supposed to pay. Can you check online the fees for this master programme? Please start your research from the homepage of the university.	Cued RTAP

9.5 Appendix E: An Overview with the Four Used Tasks with Description and Required Steps to Complete the Tasks

Task	Task description	Task steps
1	Peter, an old friend of yours, would like to follow the Master programme Health Sciences from the University of Twente. He cannot momentarily access to the Internet and he asked you to help him by finding out if there is a course that would prepare him to deal with data in the Master programme Health Sciences. Can you help him? Please start your research from the homepage of the university.	1. Start at https://www.utwente.nl/en/ 2. In the left hand navigation, click on "Education" 3. In the left hand navigation, click on "Master" 4. In the left hand navigation, click on "All Master's programmes" 5. Scroll in the text to "HEALTH SCIENCES" and click on it 6. In the left hand navigation, click on "Programme & specializations" 7. In the left hand navigation, click on "Study overview" 8. In the text, scroll to the header "GENERAL COURSES" 9. Task is succeeded if the button for the fold menu is found.
2	Francis, a friend of yours, recently finished his Bachelor, and he is thinking about a Master at the University of Twente. Since you already study at the University of Twente, he has asked you for help. Your friend was wondering if there is a specialization on the maintenance and the operations of mechanical objects that will enable him to graduate as an engineer. Can you help him? Please start your research from the homepage of the university.	1. Start at https://www.utwente.nl/en/ 2. In the left hand navigation, click on "Education" 3. In the left hand navigation, click on "Master" 4. In the left hand navigation, click on "All Master's programmes" 5. Scroll in the text to "Mechanical Engineering" 6. Click on the button "View 5 specializations" 7. Scroll to the specialization "MAINTENANCE ENGINEERING & OPERATIONS" 8. Click on the button of the specialization "MAINTENANCE ENGINEERING & OPERATIONS" 9. If this specialization is found, the task is succeeded.
3	Paula, a friend of yours, has finished her Bachelor Psychology at the University of Twente some time ago. She is interested in doing a Master programme in Systems and	1. Start at https://www.utwente.nl/en/ 2. In the left hand navigation, click on "Education"

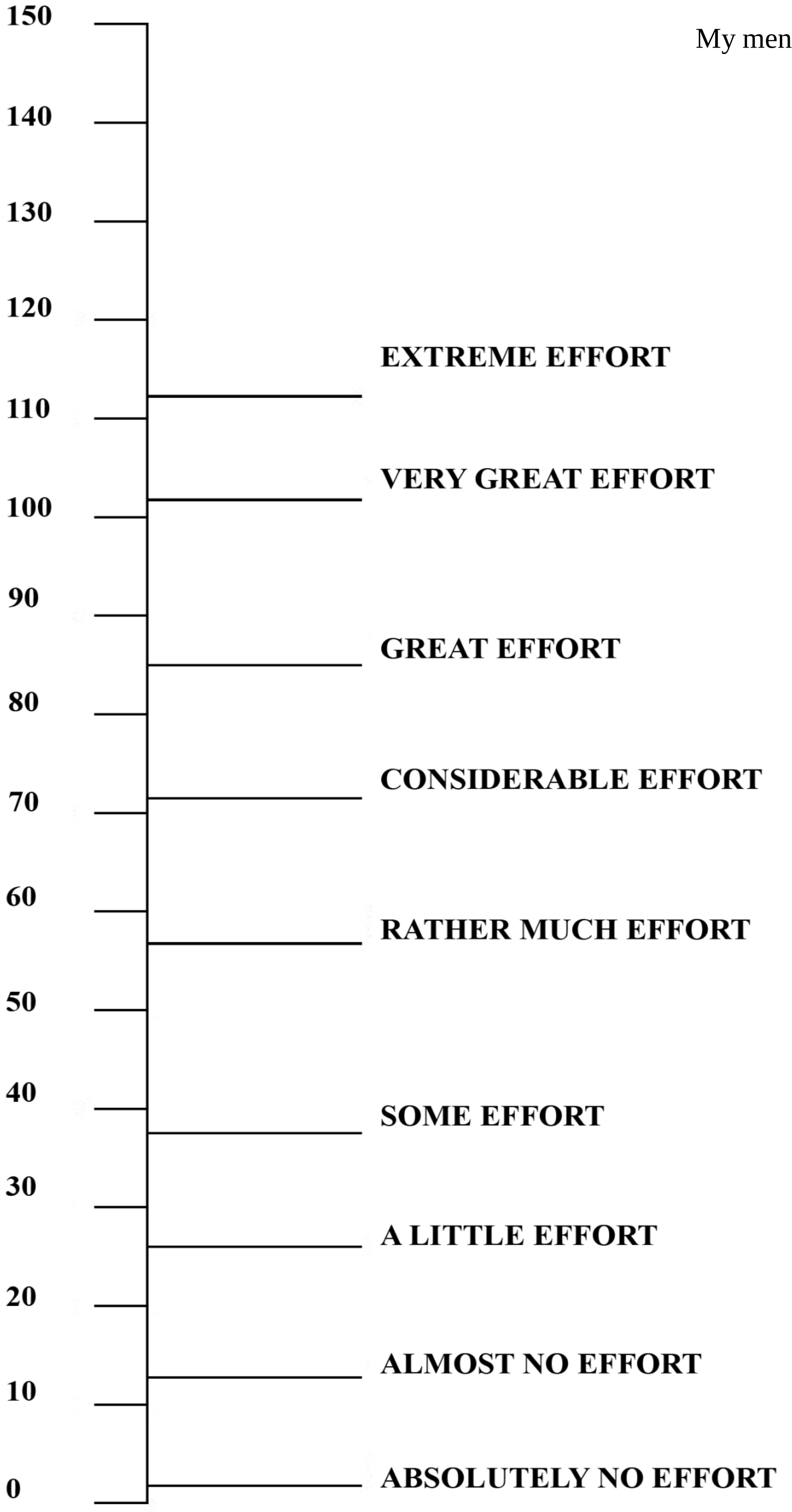
	Control at the University of Twente. However, she is not sure if she can enter this Master programme with a Bachelor in Psychology. Can you help her? Please start your research from the homepage of the university.	3. In the left hand navigation, click on “Master” 4. In the left hand navigation, click on “All Master’s programmes” 5. Scroll in the text to “SYSTEMS AND CONTROL” and click on it 6. In the left hand navigation, click on “Programme Structure” 7. In the left hand navigation, click on “Factsheet” 8. Scroll towards the header “ADMISSION REQUIREMENTS” 9. The task is succeeded if the participant cannot find the Bachelor Psychology in the lists of academic degree.
4	Ni, a Chinese friend of yours, is looking to subscribe to the Master programme Health Sciences at Twente. He would like to attend the course in September 2020. However, he is not sure how much tuition fees he is supposed to pay. Can you check online the fees for this Master Programme? Please start your research from the homepage of the university.	1. Start at https://www.utwente.nl/en/ 2. In the left hand navigation, click on “Education” 3. In the left hand navigation, click on “Master” 4. In the left hand navigation, click on “All Master’s programmes” 5. Scroll in the text to “HEALTH SCIENCES” and click on it 6. In the left hand navigation, click on “Programme & specializations” 7. In the left hand navigation, click on “Study overview” 8. Scroll towards the header “TUITION FEES 2020 / 2021” 9. If the non-EU/EER tariff of €12.250 is found, the task is succeeded.

9.6 Appendix F: Questionnaire UMUX-lite

	Strongly disagree						Strongly agree
1. The system's (website) capabilities meet my requirements.	1	2	3	4	5	6	7
2. The system (website) is easy to use.	1	2	3	4	5	6	7

9.7 Appendix G: Instrument Rating Scale Mental Effort

My mental effort was: ______



9.8 Appendix H: Randomisation of Methods and Tasks with Explanation

To randomise each technique onto each participant, the same website as before is used, namely the website random.org. Each technique is given a number between 1 and 4, to ensure that the randomisation works. By using the ‘Integer Generator’, a number between 1 and 4 will be generated. For the first participant, the ‘Integer Generator’ generated a 3, which means that the first participant got the technique ‘CTAP Bubble’, which is the concurrent think-aloud protocol with the bubble that cues the participant. This is also done for the next participants, until one of the techniques has occurred five times. Then the remaining three techniques will be given a number between 1 and 3, and it the ‘Integer Generator’ is used another time. But this time it will only be used for a number generated between 1 and 3, and it will be used for the remaining participants. This continues until the second technique has occurred five times. Then the remaining two techniques will be given the numbers 1 and 2, and as before the ‘Integer Generator’ is used. Again as before, it will only be used for a number generated that is either 1 or 2, and will be used for the remaining participants. This continues until the third technique has occurred five times. Then one or multiple participants should have not yet received an assigned technique, and therefore will receive the assignment of the technique that has yet to occur five times. In this case the nineteenth and twentieth participants got the retrospective think-aloud protocol without any additional cues as technique. The next step is to determine the order of tasks. In the first study it appeared that task 1 and task 4 are quite similar, as will be further explained in the results section. Therefore, to diminish the bias between the two tasks, it is crucial that both tasks occur in the order as either the first or the last task. Thus to determine the order of the task for each participant, the first step is determine what the first task, and thus the last task, will be. A coin flip is used to determine the first and last task. In this coin flip, for every participant the head stands for task 1 and tails stands for task 4. For example, for participant 1 the first coin flip gave tails. That means that participant 1 starts with task 4. To determine the order for task 2 and task 3, another coin flip is used. In this coin flip, head stands for task 2 and tails for task 3. Thus for example in the second coin flip for participant 1, the results was tails. That determines that full order of the four tasks, since task 2 and task 1 are also determined by the first and second coin flip. This gives the full table of randomisation to do the experiments.

Participant	Method	Order of tasks
1	CTAP Bubble	4-3-2-1
2	CTAP	1-2-3-4
3	RTAP Bubble	4-3-2-1
4	RTAP	4-3-2-1
5	RTAP	1-2-3-4
6	CTAP	1-2-3-4
7	CTAP Bubble	1-3-2-4
8	RTAP Bubble	1-3-2-4
9	RTAP Bubble	4-3-2-1
10	RTAP Bubble	1-3-2-4
11	RTAP Bubble	1-3-2-4
12	CTAP	1-2-3-4
13	CTAP Bubble	1-2-3-4
14	CTAP Bubble	4-3-2-1
15	RTAP	1-2-3-4
16	CTAP	4-3-2-1
17	CTAP Bubble	1-3-2-4
18	CTAP	4-3-2-1
19	RTAP	4-2-3-1
20	RTAP	4-2-3-1

9.9 Appendix I: The Difference-Tables from the Six Items with Significant Differences

The Difference-Table of the Item 'I Found It Unnatural to Verbalise my Thoughts' from the Participants' Experience Questionnaire with the Condition Cued RTAP as Reference Condition

fixef	center	lower	upper
Cued RTAP	0.5932119	-2.135486	4.1966638
Classic CTAP	-2.9158101	-6.343655	-0.3038166
Classic RTAP	-0.3391063	-2.557258	1.7246824
Cued CTAP	-0.5651935	-2.652142	1.3497802

The Difference-Table of the Item 'I Found It Unpleasant to Verbalise my Thoughts' from the Participants' Experience Questionnaire with the Condition Classic CTAP as Reference Condition

fixef	center	lower	upper
Classic CTAP	2.3093978	0.2164120	4.6945740
Classic RTAP	0.2987967	-1.7176091	2.3986729
Cued CTAP	2.3564932	0.2049753	4.8669362
Cued RTAP	1.7814827	-0.2282429	4.1911590

The Difference-Table of the Item 'I Found it Tiring to Verbalise my Thoughts' from the Participants' Experience Questionnaire with the condition Classic CTAP as Reference Condition

fixef	center	lower	upper
Classic CTAP	3.9802919	1.4158876	7.1541172
Classic RTAP	1.0615412	-0.9919122	3.2644831
Cued CTAP	3.8166280	1.3437559	6.8572721
Cued RTAP	3.5243837	1.1734659	6.5457795

The Difference-Table of the Item 'I Found it Tiring to Verbalise my Thoughts' from the Participants' Experience Questionnaire with the condition Classic RTAP as Reference Condition

fixef	center	lower	upper
Classic RTAP	2.9761037	0.5212300	5.7316570
Cued CTAP	2.8207805	0.5365905	5.6286322
Cued RTAP	2.5293990	0.2447469	5.2504695
Classic CTAP	-1.0436567	-3.3153037	0.9908425

Tthe Difference-Table of the Item 'I Found it Time-Consuming to Verbalise my Thoughts' from the Participants' Experience Questionnaire with the Condition Classic CTAP as Reference Condition

fixef	center	lower	upper
Classic CTAP	8.1973982	4.4897184	13.667251
Classic RTAP	2.9216344	0.3389928	6.585777
Cued CTAP	6.5138572	3.2097675	11.568087
Cued RTAP	4.8280297	1.9338880	8.995432

The Difference-Table of the Item 'I Found it Time-Consuming to Verbalise my Thoughts' from the Participants' Experience Questionnaire *with the condition Classic RTAP as Reference Condition*

fixef	center	lower	upper
Classic RTAP	5.2541041	2.0491928	9.7132083
Cued CTAP	3.5215702	0.9650025	7.2771565
Cued RTAP	1.8395397	-0.3130212	4.5289227
Classic CTAP	-2.9929410	-6.6472183	-0.3481863

The Difference-Table of the Item 'The Links in the Texts Help me to Easily Find more Information on Specific Subjects' from the Participants' Experience Questionnaire with the Condition Cued RTAP as Reference Condition

fixef	center	lower	upper
Cued RTAP	5.4190113	2.0351117	9.9167818
Classic CTAP	2.0154010	-0.2354895	4.6618600
Classic RTAP	1.9060153	-0.4159091	4.7752999
Cued CTAP	4.3464974	1.6758175	7.7505794

The Difference-Table of the Item 'The Content on the Master Programme Page Persuades me *to Read more about the Programme' from the Participants' Experience Questionnaire with the Condition Classic RTAP as Reference Condition*

fixef	center	lower	upper
Classic RTAP	-0.5200473	-2.599637	1.3511405
Cued CTAP	-5.4864012	-10.153833	-2.0249114
Cued RTAP	-2.5835314	-6.318517	0.2473876
Classic CTAP	-3.7535882	-8.097477	-0.6258866

9.10 Appendix J: R Script for the Analysis

Loading the packages from R that can be useful in this analysis.

```r
library(tidyverse)
library(knitr)
library(rstanarm)
library(mascutils)
library(brms)
library(bayr)
```

All the Excel files are loaded that are going to be used in this analysis.

```r
descriptive <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Descriptive.csv")

Part_exp <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Part Exp All.csv")

Performance <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Performance.csv")

RSME <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/RSME.csv")

UMUX <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UMUX-Lite.csv")

UP <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UP.csv")
```

The first step is the analysis of the descriptive statistics from the sample of participants. So the data from the excel files is loaded again.

```r
descriptive <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Descriptive.csv")
```

The next step is to create a summary of the descriptive data, to use later in the methods section.

```r
summary(descriptive)
```

The next step is the analysis of the performance from the participants. Four data sets are loaded from Excel files, to have a different intercept later in the analysis for the four conditions of the current study

```
Performance1 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study
2/Performance1.csv")

Performance2 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study
2/Performance2.csv")

Performance3 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study
2/Performance3.csv")

Performance4 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study
2/Performance4.csv")
```

The next step is to do the regression analyses. Four different analyses are done with four different intercepts. The family is exGaussian, because the data cannot be zero. The function 'iter' is changed from its 2000 to 15000, because more iteration is necessary to complete the analyses.

```
m_1 <- brm(ToT ~ condition, family = exgaussian(), data = Performance1, ite
r = 15000)
fixef(m_1)

m_2 <- brm(ToT ~ condition, family = exgaussian(), data = Performance2, ite
r = 15000)
fixef(m_2)

m_3 <- brm(ToT ~ condition, family = exgaussian(), data = Performance3, ite
r = 15000)
fixef(m_3)

m_4 <- brm(ToT ~ condition, family = exgaussian(), data = Performance4, ite
r = 15000)
fixef(m_4)
```

Since the performance consists of both ToT and the completion of the tasks, the next step is to look at the completion and do the regression analyses. Four different analyses are done with four different intercepts. The family is binomial, because the count data has a clear upper limit.

```
m_5 <- brm(complete ~ condition, family = binomial(), data = Performance1)
fixef(m_5)

m_6 <- brm(complete ~ condition, family = binomial(), data = Performance2)
fixef(m_6)

m_7 <- brm(complete ~ condition, family = binomial(), data = Performance3)
fixef(m_7)

m_8 <- brm(complete ~ condition, family = binomial(), data = Performance4)
fixef(m_8)
```

The next step is to analyse the Rating Scale Mental Effort (RSME). Four data sets are loaded from Excel files, to have a different intercept later in the analysis for the four conditions of the current study

```
RSME1 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/RSME1.csv")
RSME2 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/RSME2.csv")
RSME3 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/RSME3.csv")
RSME4 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/RSME4.csv")
```

The next step is to do the regression analyses. Four different analyses are done with four different intercepts. The family is binomial, because the data has a clear upper limit.

```
m_9 <- brm(RSME ~ condition, family = binomial(), data = RSME1)
fixef(m_9)

m_10 <- brm(RSME ~ condition, family = binomial(), data = RSME2)
fixef(m_10)

m_11 <- brm(RSME ~ condition, family = binomial(), data = RSME3)
fixef(m_11)

m_12 <- brm(RSME ~ condition, family = binomial(), data = RSME4)
fixef(m_12)
```

The next step is to analyse the UMUX-lite. Four data sets are loaded from Excel files, to have a different intercept later in the analysis for the four conditions of the current study

```
UMUX1 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UMUX-Lite1.cs
v")
UMUX2 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UMUX-Lite2.cs
v")
UMUX3 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UMUX-Lite3.cs
v")
UMUX4 <-
```

```
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UMUX-Lite4.cs
v")
```

To get the mean UMUX-lite score per condition, four different data sets are created. Then the function summary gives the mean per condition. After that, it is time to do the regression analyses. Four different analyses are done with four different intercepts. The family is binomial, because the data has a clear upper limit.

```
UMUXCLC <-
  UMUX %>%
  filter(condition == "Classic CTAP")

UMUXCLR <-
  UMUX %>%
  filter(condition == "Classic RTAP")

UMUXCUC <-
  UMUX %>%
  filter(condition == "Cued CTAP")

UMUXCUR <-
  UMUX %>%
  filter(condition == "Cued RTAP")

summary(UMUXCLC)

summary(UMUXCLR)

summary(UMUXCUC)

summary(UMUXCUR)

m_13 <- brm(Score ~ condition, family = binomial(), data = UMUX1)
fixef(m_13)

m_14 <- brm(Score ~ condition, family = binomial(), data = UMUX2)
fixef(m_14)

m_15 <- brm(Score ~ condition, family = binomial(), data = UMUX3)
fixef(m_15)

m_16 <- brm(Score ~ condition, family = binomial(), data = UMUX4)
fixef(m_16)
```

The next step is to analyse the questionnaires from the participants' experience. Four data sets are loaded from Excel files, to have a different intercept later in the analysis for the four conditions of the current study

```
Part_exp1 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Part Exp All1
.csv")
Part_exp2 <-
```

```
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Part Exp All2
.csv")
Part_exp3 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Part Exp All3
.csv")
Part_exp4 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/Part Exp All4
.csv")
```

Then, four different filters are applied because of missing scores by some participants. This is because different participants followed different conditions and each condition had a different questionnaire. The two conditions with the retrospective think aloud protocol had questions about the video that the participants saw after they were done with the task. And the two cued conditions had questions about the bubble that cued the participants. Therefore each condition had a different questionnaire.

```
PE1 <-
  Part_exp %>%
  filter(!is.na(video.distracted))

PE2 <-
  Part_exp %>%
  filter(!is.na(video.disliked))

PE3 <-
  Part_exp %>%
  filter(!is.na(bubble.distracted))

PE4 <-
  Part_exp %>%
  filter(!is.na(bubble.disliked))
```

The next step is to analyse each question from every questionnaire that test the experience from the participants by regression analysis. Four different analyses are done per question with four different intercepts, although there are some exceptions. Some questions have to do with two intercepts, thus one analysis is sufficient. The family that is used in these analyses is cratio, and is commonly used for rating scales.

Some analyses differ from the rest, because of the added functions ‘iter’ and ‘control’. The function ‘iter’ is changed from its 2000 to 15000, because more iteration is necessary to complete the analyses. And the function ‘control’ is added to diminish the bias in obtained posterior samples.

I found it difficult to verbalise my thoughts.

```
m_17 <- brm(difficult.verbalise ~ condition, family = cratio(), data = Part
_exp1)
fixef(m_17)
```

```r
m_18 <- brm(difficult.verbalise ~ condition, family = cratio(), data = Part
_exp2)
fixef(m_18)

m_19 <- brm(difficult.verbalise ~ condition, family = cratio(), data = Part
_exp3)
fixef(m_19)

m_20 <- brm(difficult.verbalise ~ condition, family = cratio(), data = Part
_exp4)
fixef(m_20)
```

I found it unnatural to verbalise my thoughts.

```r
m_21 <- brm(unnatural.verbalise ~ condition, family = cratio(), data = Part
_exp1)
fixef(m_21)

m_22 <- brm(unnatural.verbalise ~ condition, family = cratio(), data = Part
_exp2)
fixef(m_22)

m_23 <- brm(unnatural.verbalise ~ condition, family = cratio(), data = Part
_exp3)
fixef(m_23)

m_24 <- brm(unnatural.verbalise ~ condition, family = cratio(), data = Part
_exp4)
fixef(m_24)
```

I found it unpleasant to verbalise my thoughts.

```r
m_25 <- brm(unpleasant.verbalise ~ condition, family = cratio(), data = Par
t_exp1)
fixef(m_25)

m_26 <- brm(unpleasant.verbalise ~ condition, family = cratio(), data = Par
t_exp2)
fixef(m_26)

m_27 <- brm(unpleasant.verbalise ~ condition, family = cratio(), data = Par
t_exp3)
fixef(m_27)

m_28 <- brm(unpleasant.verbalise ~ condition, family = cratio(), data = Par
t_exp4)
fixef(m_28)
```

I found it tiring to verbalise my thoughts.

```r
m_29 <- brm(tiring.verbalise ~ condition, family = cratio(), data = Part_ex
p1)
fixef(m_29)
```

```
m_30 <- brm(tiring.verbalise ~ condition, family = cratio(), data = Part_ex
p2)
fixef(m_30)

m_31 <- brm(tiring.verbalise ~ condition, family = cratio(), data = Part_ex
p3)
fixef(m_31)

m_32 <- brm(tiring.verbalise ~ condition, family = cratio(), data = Part_ex
p4)
fixef(m_32)
```

I found it time-consuming to verbalise my thoughts.

```
m_33 <- brm(timeconsuming.verbalise ~ condition, family = cratio(), data =
Part_exp1)
fixef(m_33)

m_34 <- brm(timeconsuming.verbalise ~ condition, family = cratio(), data =
Part_exp2)
fixef(m_34)

m_35 <- brm(timeconsuming.verbalise ~ condition, family = cratio(), data =
Part_exp3)
fixef(m_35)

m_36 <- brm(timeconsuming.verbalise ~ condition, family = cratio(), data =
Part_exp4)
fixef(m_36)
```

Seeing the gaze-bubble distracted me in remembering what I thought.

```
m_37 <- brm(bubble.distracted ~ condition, family = cratio(), data = PE3)
fixef(m_37)
```

I disliked seeing the gaze-bubble while verbalising my thoughts.

```
m_38 <- brm(bubble.disliked ~ condition, family = cratio(), data = PE4)
fixef(m_38)
```

Seeing the playback video distracted me in remembering what I thought.

```
m_39 <- brm(video.distracted ~ condition, family = cratio(), data = PE1)
fixef(m_39)
```

I disliked seeing the playback video while verbalising my thoughts.

```
m_40 <- brm(video.disliked ~ condition, family = cratio(), data = PE2)
fixef(m_40)
```

The presence of the researcher was unnatural.

```r
m_41 <- brm(presence.unnatural ~ condition, family = cratio(), data = Part_
exp1, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_41)

m_42 <- brm(presence.unnatural ~ condition, family = cratio(), data = Part_
exp2, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_42)

m_43 <- brm(presence.unnatural ~ condition, family = cratio(), data = Part_
exp3, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_43)

m_44 <- brm(presence.unnatural ~ condition, family = cratio(), data = Part_
exp4, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_44)
```

The presence of the researcher was disturbing.

```r
m_45 <- brm(presence.disturbing ~ condition, family = cratio(), data = Part
_exp1, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_45)

m_46 <- brm(presence.disturbing ~ condition, family = cratio(), data = Part
_exp2, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_46)

m_47 <- brm(presence.disturbing ~ condition, family = cratio(), data = Part
_exp3, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_47)

m_48 <- brm(presence.disturbing ~ condition, family = cratio(), data = Part
_exp4, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_48)
```

The presence of the researcher was unpleasant.

```r
m_49 <- brm(presence.unpleasant ~ condition, family = cratio(), data = Part
_exp1, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_49)

m_50 <- brm(presence.unpleasant ~ condition, family = cratio(), data = Part
_exp2, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_50)

m_51 <- brm(presence.unpleasant ~ condition, family = cratio(), data = Part
_exp3, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_51)

m_52 <- brm(presence.unpleasant ~ condition, family = cratio(), data = Part
_exp4, iter = 15000, control = list(adapt_delta = 0.99))
fixef(m_52)
```

The links in the texts help me to easily find more information on specific subjects.

```
m_53 <- brm(links ~ condition, family = cratio(), data = Part_exp1)
fixef(m_53)

m_54 <- brm(links ~ condition, family = cratio(), data = Part_exp2)
fixef(m_54)

m_55 <- brm(links ~ condition, family = cratio(), data = Part_exp3)
fixef(m_55)

m_56 <- brm(links ~ condition, family = cratio(), data = Part_exp4)
fixef(m_56)
```

The content on the Master Programme page persuades me to read more about the programme.

```
m_57 <- brm(content ~ condition, family = cratio(), data = Part_exp1)
fixef(m_57)

m_58 <- brm(content ~ condition, family = cratio(), data = Part_exp2)
fixef(m_58)

m_59 <- brm(content ~ condition, family = cratio(), data = Part_exp3)
fixef(m_59)

m_60 <- brm(content ~ condition, family = cratio(), data = Part_exp4)
fixef(m_60)
```

I like the tone of voice that is used in the Master Programme site.

```
m_61 <- brm(tone.of.voice ~ condition, family = cratio(), data = Part_exp1)
fixef(m_61)

m_62 <- brm(tone.of.voice ~ condition, family = cratio(), data = Part_exp2)
fixef(m_62)

m_63 <- brm(tone.of.voice ~ condition, family = cratio(), data = Part_exp3)
fixef(m_63)

m_64 <- brm(tone.of.voice ~ condition, family = cratio(), data = Part_exp4)
fixef(m_64)
```

The next step is to analyse the usability problems. Four data sets are loaded from Excel files, to have a different intercept later in the analysis for the four conditions of the current study

```
UP1 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UP1.csv")
UP2 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UP2.csv")
UP3 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UP3.csv")
UP4 <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/UP4.csv")
```

The next step is to analyse the amount of usability problems with regression analyses. Four different analyses are done with four different intercepts. The family that is used in these analyses is poisson, because there is no apparent upper limit.

Besides the amount of usability problems, the unique usability problems are also analysed. This is also done with four different analyses. The family that is used in the analyses with unique usability problems is also poisson, because there is no apparent upper limit. These analyses also differ, because of the added functions ‘iter’ and ‘control’. The function ‘iter’ is changed from its 2000 to 15000, because more iteration is necessary to complete the analyses. And the function ‘control’ is added to diminish the bias in obtained posterior samples.

```
m_65 <- brm(Total ~ Condition, family = poisson(), data = UP1)
fixef(m_65)

m_66 <- brm(Total ~ Condition, family = poisson(), data = UP2)
fixef(m_66)

m_67 <- brm(Total ~ Condition, family = poisson(), data = UP3)
fixef(m_67)

m_68 <- brm(Total ~ Condition, family = poisson(), data = UP4)
fixef(m_68)


m_69 <- brm(TU ~ Condition, family = poisson(), data = UP1, iter = 15000, c
ontrol = list(adapt_delta = 0.99, max_treedepth = 15))
fixef(m_69)

m_70 <- brm(TU ~ Condition, family = poisson(), data = UP2, iter = 15000, c
ontrol = list(adapt_delta = 0.99, max_treedepth = 15))
fixef(m_70)

m_71 <- brm(TU ~ Condition, family = poisson(), data = UP3, iter = 15000, c
ontrol = list(adapt_delta = 0.99, max_treedepth = 15))
fixef(m_71)

m_72 <- brm(TU ~ Condition, family = poisson(), data = UP4, iter = 15000, c
ontrol = list(adapt_delta = 0.99, max_treedepth = 15))
fixef(m_72)
```

After analysing the amount of usability problems, the next steps are to analyse the severity level, the method the usability problems are detected, and what type the usability problems are. Four different analyses are done with four different intercepts.

Severity levels

For the severity levels, the family that is used is cratio, since it is commonly used for rating scales.

```r
m_73 <- brm(Severity ~ Condition, family = cratio(), data = UP1)
fixef(m_73)

m_74 <- brm(Severity ~ Condition, family = cratio(), data = UP2)
fixef(m_74)

m_75 <- brm(Severity ~ Condition, family = cratio(), data = UP3)
fixef(m_75)

m_76 <- brm(Severity ~ Condition, family = cratio(), data = UP4)
fixef(m_76)
```

Detection methods

For the detection methods, the family that is used is acat, which is used with adjacent categories.

```r
m_77 <- brm(Detection ~ Condition, family = acat(), data = UP1)
fixef(m_77)

m_78 <- brm(Detection ~ Condition, family = acat(), data = UP2)
fixef(m_78)

m_79 <- brm(Detection ~ Condition, family = acat(), data = UP3)
fixef(m_79)

m_80 <- brm(Detection ~ Condition, family = acat(), data = UP4)
fixef(m_80)
```

Types of usability problems

To analyse the types of usability problems, a filter is applied four times to the four data sets with different intercepts. The four different types of usability problems are isolated from each other to analyse each type individually. The family that is used is poisson, since the amounts have no clear upper limit.

```r
UP1T <-
  UP1 %>%
  filter(Type == "1")

UP2T <-
  UP2 %>%
  filter(Type == "1")

UP3T <-
  UP3 %>%
  filter(Type == "1")
```

```r
UP4T <-
  UP4 %>%
  filter(Type == "1")

UP1N <-
  UP1 %>%
  filter(Type == "2")

UP2N <-
  UP2 %>%
  filter(Type == "2")

UP3N <-
  UP3 %>%
  filter(Type == "2")

UP1D <-
  UP1 %>%
  filter(Type == "3")

UP2D <-
  UP2 %>%
  filter(Type == "3")

UP3D <-
  UP3 %>%
  filter(Type == "3")

UP4D <-
  UP4 %>%
  filter(Type == "3")

UP1C <-
  UP1 %>%
  filter(Type == "4")

UP2C <-
  UP2 %>%
  filter(Type == "4")

UP3C <-
  UP3 %>%
  filter(Type == "4")

UP4C <-
  UP4 %>%
  filter(Type == "4")

m_81 <- brm(Type ~ Condition, family = poisson(), data = UP1T)
fixef(m_81)

m_82 <- brm(Type ~ Condition, family = poisson(), data = UP2T)
fixef(m_82)
```

```r
m_83 <- brm(Type ~ Condition, family = poisson(), data = UP3T)
fixef(m_83)

m_84 <- brm(Type ~ Condition, family = poisson(), data = UP4T)
fixef(m_84)

m_85 <- brm(Type ~ Condition, family = poisson(), data = UP1N)
fixef(m_85)

m_86 <- brm(Type ~ Condition, family = poisson(), data = UP2N)
fixef(m_86)

m_87 <- brm(Type ~ Condition, family = poisson(), data = UP3N)
fixef(m_87)

m_88 <- brm(Type ~ Condition, family = poisson(), data = UP1D)
fixef(m_88)

m_89 <- brm(Type ~ Condition, family = poisson(), data = UP2D)
fixef(m_89)

m_90 <- brm(Type ~ Condition, family = poisson(), data = UP3D)
fixef(m_90)

m_91 <- brm(Type ~ Condition, family = poisson(), data = UP4D)
fixef(m_91)

m_92 <- brm(Type ~ Condition, family = poisson(), data = UP1C)
fixef(m_92)

m_93 <- brm(Type ~ Condition, family = poisson(), data = UP2C)
fixef(m_93)

m_94 <- brm(Type ~ Condition, family = poisson(), data = UP3C)
fixef(m_94)

m_95 <- brm(Type ~ Condition, family = poisson(), data = UP4C)
fixef(m_95)
```

The last step of the analysis is the analysis of the results from the Tobii Pro Lab. First, the data from the Excel files are loaded again.

```r
TobiiFC <-
  read.csv("D:/Arjan/Documents/R studio/Master/Thesis/Study 2/TobiiFC.csv")
```

The next step is to analyse the difference between amounts of fixations per condition. As mentioned, the two conditions classic CTAP and classic RTAP have generated fixation data. Therefore only these two conditions will be analysed. The family that will be used is poisson, because this data has no clear upper limit.

```r
m_117 <- brm(Sum ~ Condition, family = poisson(), data = TobiiFC)
fixef(m_117)
```

The next step is to compare the count of fixations with the usability problems that were verbalised by the participants, to see what the effect is on both. First a filter is applied to select the verbalised usability problems from the two classic conditions.

```
UPV <-
  UP %>%
  filter(Detection == "V") %>%
  filter(!Condition == "Cued CTAP") %>%
  filter(!Condition == "Cued RTAP")
```

Then the regression analysis is done on the count of fixations with usability problems that were verbalized by the participants. The family that is used is poisson for there is no clear upper limit.

```
m_118 <- brm(Total ~ Condition, family = poisson(), data = UPV)
fixef(m_118)
```