Gottfried Wilhelm
Leibniz Universität Hannover
Faculty of Electrical Engineering and Computer Science
Institute of Practical Computer Science
Software Engineering Group

Using Eye-Tracking to Detect Explainability Needs

Bachelor's Thesis

in Computer Science

by

Karl Lasse Borcherding

First Examiner: Prof. Dr. Kurt Schneider
Second Examiner: Dr. Jil Klünder
Supervisor: Jakob Droste, M. Sc.

Hannover, 11.07.2024

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelor's Thesis selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 11.07.2024

Karl Lasse Borcherding

Abstract

Software is becoming increasingly complex, which increases the relevance of explainability. Usability guidelines often help to ensure the usability of complex systems. However, usability without explanations has its limits at some point. Efficient methods are therefore needed to integrate explainability into software without compromising usability.

Eye trackers offer the possibility of analysing eye movements in detail. When there is a need for explanation, a user actively searches for information in the system, which can manifest itself in characteristic eye movements. If an eye tracker recognises this behaviour, developers can react in a precise way to efficiently satisfy the need for explanation.

In this work, a study was conducted with eleven participants who had to complete three tests designed to generate a need for explanation. During the tests, eye movements were recorded using an eye tracker. After each test, the participants stated their need for explanation, their stress level and additional informations.

The analysis of the study provides valuable insights about how the need for explanation can be identified on the basis of eye movements. In particular, it was shown that there is a significant correlation between eye movements and the need for explanation at an individual level. In addition, the eye tracker was able to record stress and identify preferred areas of fixation that were increasingly focused by the participants when there was a need for explanation. A generalisable correlation between eye movements and the need for explanation could not be established.

Zusammenfassung

Software wird zunehmend komplexer, was die Relevanz von Erklärbarkeit steigert. Usability-Richtlinien helfen oft, die Bedienbarkeit komplexer Systeme zu gewährleisten. Allerdings stößt Usability ohne Erklärungen irgendwann an ihre Grenzen. Daher sind effiziente Methoden notwendig, um Explainability in Software zu integrieren, ohne die Usability zu beeinträchtigen.

Eyetracker bieten die Möglichkeit, Augenbewegungen detailliert zu analysieren. Bei Erklärungsbedarf sucht ein Nutzer aktiv nach Informationen im System, was sich in charakteristischen Augenbewegungen äußern kann. Wenn ein Eyetracker dieses Verhalten erkennt, können Entwickler gezielte Maßnahmen ergreifen, um den Erklärungsbedarf effizient zu decken.

In dieser Arbeit wurde eine Studie mit elf Teilnehmern durchgeführt, die drei Tests absolvieren mussten, um Erklärungsbedarf zu erzeugen. Während der Tests wurden die Augenbewegungen mittels Eyetracker aufgezeichnet. Nach jedem Test gaben die Probanden ihren Erklärungsbedarf, ihr Stresslevel und zusätzliche Informationen an.

Die Auswertung der Studie liefert wertvolle Erkenntnisse darüber, wie sich Erklärungsbedarf anhand von Augenbewegungen identifizieren lässt. Es zeigte sich insbesondere, dass auf individueller Ebene eine signifikante Korrelation zwischen Augenbewegungen und Erklärungsbedarf besteht. Zudem konnte der Eyetracker Stress erfassen und bevorzugte Fixationsbereiche identifizieren, die von den Probanden bei Erklärungsbedarf verstärkt fokussiert wurden. Ein generalisierbares Korrelationsverhältnis zwischen den Augenbewegungen und dem Erklärungsbedarf konnte nicht festgestellt werden.

Contents

Chapter 1

Introduction

1.1 Motivation

Software systems are constantly growing in complexity, which is why the relevance of explainability is also constantly increasing [20]. Despite the use of the user's existing knowledge base and specific usability guideliness, such as the use of the same symbols for the same functions or the integration of the same hotkeys for similar and identical methods, the amount of possible interactions and system behaviors in some software is so enormous that it seems impossible to operate software in its entirety without explanations [6]. In addition to the frustration that could arise for the user due to difficulties in using the software, trying to explain too much to the user can also have a counterproductive effect [9]. An overloaded explanation, or an explanation that appears at the wrong moment, does not help the user and damages the user's trust in the software [28]. To counteract this problem, the use of eye trackers could be helpful. Eye trackers can determine the user's fixations and saccades very precisely [2, 29]. This could give developers the opportunity to incorporate targeted explanations exactly where they are needed by the user.

1.2 Problem Statement

As just described in the above section 1.1, the availability of the right explanation at the right time for the user is important to keep the frustration level low when the need for an explanation arises [28]. It is now interesting to see whether and how software can be designed to always display an explanation at the right moment when the user might need it. It should also avoid giving the user an unwanted explanation when there is no need for one. The question that therefore arises is when the user needs an explanation and whether this can be determined automatically. Whether the data that can be collected is accurate enough to be used for such software also remains to be seen. In the context of this scientific work, attention is primarily paid to the data from eye trackers.

1.3 Solution Approach

How exactly a user behaves when they need an explanation differs from user to user [15]. However, common behaviours between users could be found using the data from an eye tracker when an explanation is needed. As part of a study, 11 test subjects had to complete various software-related tasks under time pressure. The tasks were designed to ensure that the test subjects were highly unlikely to be able to complete some tasks without guidance. The data that the eye tracker then outputs for tasks where the user's needed explanation is then compared with data where the user did not need any explanation to complete the tasks. In order to validate the data, the test subjects were also given a small questionnaire for each individual test, on which they were asked to indicate their level of need for explanation and stress together with some other data. By comparing the above data, it is possible to determine whether the eye tracker can perceive and measure the need for explanation.

1.4 Thesis Structure

This thesis is structured as follows. The "Background and Related Work" chapter 2 presents the background information that is necessary for understanding this scientific work. First, some aspects of explainability and its effects in software development are explained. This is followed by an excursion into eye trackers and study design. At the end of the chapter, the related work of this thesis is also presented. In the chapter "Research Design" 3, two research questions are defined and an insight into the study design and data analysis is given. The "Evaluation" chapter 4 presents the analysed data from the study using diagrams with associated explanations. In the following chapter "Discussion" 5, the research questions are answered and the limitations of this work are mentioned. Finally, the chapter "Conclusion and Future Work"6 summarises the results of the work and provides an outlook for future work.

Chapter 2

Background and Related Work

2.1 Explainability

2.1.1 Definition

Due to the growing complexity of software, the relevance of explanations is constantly increasing to avoid overwhelming the user when using this software [9, 27]. Explainability is recognized as a non-functional requirement (NFR) and influences many other quality characteristics of software, such as usability, user experience, user efficiency and user satisfaction, both positively and negatively [8, 9].

Explainability and what makes it important can be defined in many different ways [8]. Chazette et al. [8] defines explainability as follows: "A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C". This definition is very formal and should only serve as an example, as it is too broad for this paper and not all of the variables named in it are used. A similar definition is provided by Deters et al. [12] and entails the following: "Explainability is the ability of a software to be explained to an addressee, given a specific context of use and depending on the goals of the explainer". Noteworthy about this definition is that explainability also depends on the intent of the explainer, which may vary depending on the use case. Upon closer research, however, it becomes apparent that no clear definition of explainability has yet been agreed upon in scientific discourse [25]. One thing that all definitions usually have in common is that explainability requires access to knowledge through explanations [24]. To keep it simple, this will rely on the definition by Deters et al. [12].

2.1.2 The Need for Explanation

If a user does not know how to use a software program or how to achieve their goals, they will need explanations. This need may lead to a number of problems. In addition to the user becoming frustrated, they may also lose trust in the software or the resulting outcomes [10]. Explanations can help users to understand a software to the extent that they are able to comprehend and use it [21]. However, giving the user too much or too complex information about the software can lead to a feeling of loss of control [28]. Although more information usually leads to users making better decisions, the time required to grasp the information can noticeably limit efficiency [8]. Previous research has also shown that explanations do not always offer added value and can be rather disruptive to the user in terms of the clarity and usability of a software [8]. In general, users are interested in receiving explanations in order to better understand software issues and build up a knowledge base [5]. Droste et al. [14] discovered through an online survey that one of the most common origins of explanatory needs are interaction explanations. Explanations of system behavior were also a common cause. In comparison, the need for explanations was rarely caused by incomprehensibilities in the areas of domain knowledge, privacy & security and the user interface. Whether and how this need for explanation can be measured by physiological triggers is still a relatively unexplored question in the field of explainability. Deters et al. [13] collected data in one of their studies that indicate that the need for explanation can be measured by physiological triggers.

2.2 Eyetracker

As the Tobii x3-120 eye tracker was used in the user study concucted in the context of this work, the focus here was deliberately placed on its characteristics and special features. The specific processes described also relate to the special features provided by Tobii's software and hardware. The official website of Tobii [1, 2] provides the most relevant information relating to the use of Tobii eye trackers.

2.2.1 Eyetrackers Functions and Uses

Eye trackers have long been used to study the visual attention of test subjects. One of the most widely used techniques for this is pupil center-corneal reflection (PCCR) [2]. This technique measures the reflection of light on the cornea relative to the position of the pupil centre in real time. The eye tracker's technology allows the user to move their head more freely without losing the track of the eyes. In addition, only one camera and light source are required and calibration is quicker and easier to perform [29].

2.2.2 Eyetracker made by Tobii

The figure 2.1 shows the basic functionality of eye trackers. Eye trackers work with cameras, illuminators and algorithms. The illuminators create a pattern of near-infrared light on the eyes. Cameras take high-resolution images of the eyes on which the infrared patterns are located. These images are then analysed using PCCR. Based on the data, the eye position and viewpoint are determined on the PC screen [2].

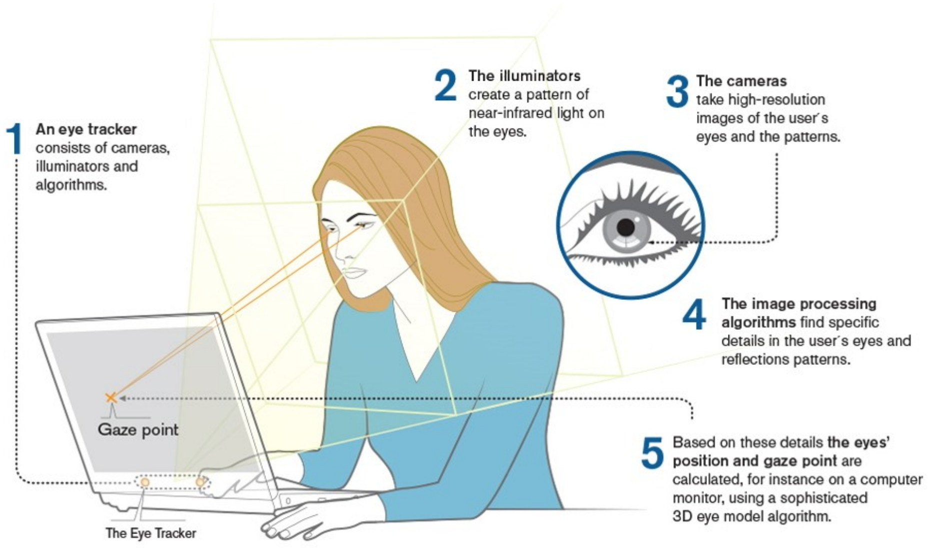

1 An eye tracker consists of cameras, illuminators and algorithms.
2 The illuminators create a pattern of near-infrared light on the eyes.
3 The cameras take high-resolution images of the user's eyes and the patterns.
4 The image processing algorithms find specific details in the user's eyes and reflections patterns.
5 Based on these details the eyes' position and gaze point are calculated, for instance on a computer monitor, using a sophisticated 3D eye model algorithm.
Gaze point
The Eye Tracker

Figure 2.1: Representation of the basic function of a screen-based eye tracker, adapted from [2]

Human eyes are slightly different by nature. The fovea, for example, sits differently in some cases, which means that the viewing angle of eyes can differ in order to look at the same thing. For calibration (the process where the hardware is adapted to the user in order to achieve the best possible results), Tobii eyetrackers use a procedure where an algorithm uses some test data (some targets on the screen that the user should stare at) to adapt the calculation of the viewing points to the user. On the Tobii website [1], the entire process is divided into the following 3 steps:

The first step is the data collection phase, where the subjects are asked to look at a few fixed objects (targets) on the screen. The number of targets may vary depending on how the experimenter has set them.

The second step is the optimisation phase, where the standard 3D eye model or the 3D eye model from the last user is adapted by Tobii to the new user for optimisation. The data from phase 1 is mapped onto the screen using the standard 3D eye model or that of the last user. The distance between the actual target and the calculated location from the mapped data is then calculated. The parameters of the 3D eye model are adjusted on the basis of this distance. With the updated model, the data from the beginning is mapped to the screen again and the distance to the calibration points is calculated. If the distances with the updated model were better than those of the previous model, the older model is discarded. This entire process is repeated until there is no further improvement in the results. The model is now considered optimised. The results of this process are shown in the calibration feedback diagram after optimisation.

The third and last step is the validation phase, where the eye model adapted to the test person is validated with new target objects. For validation, 4 new targets are used, which do not correspond in position to those from phases one and two. To collect data for validation, the subject must also fixate these new targets for a few seconds. The optimised 3D eye model is then used to map the collected data back onto the screen and the distance to the 4 target objects is measured for validation. This results in data relating to accuracy (average distance of gaze samples to validation targets), precision (standard deviation or root mean square of the distance between the gaze samples) and the number of valid eye tracking samples.

At the end, the system outputs this data together with an illustration of the calibration and validation errors and a table of the numerical results from the validation. The experimenter must then decide whether to accept or reject the data. If he accepts it, the new 3D eye model configuration is saved on the eye tracker. This configuration is used until a new calibration should take place [1].

2.3 The Art of Research Design

There are many ways to design a study, and most design ideas target specific aspects that best fit a research question [3]. For example, some types of studies are designed to investigate and answer explicit research questions, while other studies only try to observe a situation from the outside, to create a basis for more in-depth research [11, 23]. A distinction is made between qualitative and quantitative research [18]. Qualitative research is used to gain a deeper understanding of an scientific area and to understand situations, individuals or groups of individuals and their behaviour. Methods of qualitative research are: Artifact collection (for example user reviews), diary studies (for example password usage), interviews and focus groups (based on notes, transcriptions and artifacts) [16]. Quantitative research aims to prove or disprove hypotheses. This is usually done through the use of experiments that measure certain measurable values and generate analysable data [26]. Likert scales are often used, including in the study from this thesis, for example to incorporate the cognitive processes and assessments of the test subjects into the results of the experiment [22, 19]. Depending on the aim of the experiment, the collected data can be analysed differently with different focuses [19]. The study designer is allowed to be very creative in how they design a study to obtain meaningful results for the research, but the decisions made in the design of the study can strongly influence the outcome and validity of the study [4, 26].

2.4 Related Work

2.4.1 Explanations on Demand

Deters et al. [13] investigated physiological triggers for stress in order to determine the need for explanation in test subjects. In the study, the need for explanation was treated as a cause of stress. Nine study participants were equipped with biometric sensors and were asked to fulfil tests in Excel that were intended to generate explanatory needs. The data from the biometric sensors, the screen recordings from the tests and demographic data provided by the test subjects were then compared with each other. Finally, the work determined which biometric sensor was best able to recognise the need for explanation and whether this sensor is capable of being used to recognise the need for explanation.

2.4.2 Component for analysing the need for explanation

Fechner [17] designed a technique that can be used to determine the actual need for explanation. In a study with 21 participants, he had the participants operate with a prototype that provided various explanations for questions that occurred. During the study, the number of times each explanation was opened were counted. At the end, he also made statements about which types of explanation are particularly useful and in which situations explanations are particularly needed.

2.4.3 Eye Gaze as an Indicator for Stress Level

Jyotsna et al. [7] tried to determine the stress of students with the help of an eyetracker. For this purpose, two test subjects were confronted with various stimuli such as mathematical tasks, videos and questionnaires. Factors such as pupil diameter and blinking frequency were used to determine cognitive load and tiredness, which were then analysed as a factor for stress. At the end of the entire study, it was found that eye trackers can be used to determine stress very accurately and that there is a correlation between the data from the eye tracker and the stress of the test person.

Chapter 3

Research Design

3.1 Research Questions

RQ1: What data provided by an eye tracker can be used to detect irregularities in the behaviour of test subjects when explanations are required?

RQ1 is the most important and interesting research question in this thesis, which is also the most relevant in the study design process. In order to design a suitable study that can provide an answer to this question, a need for explanation must be generated among the study participants. However, in order to be able to discover particular behaviors when there is a need for explanation, comparable data is also required for situations where there is no need for explanation on the part of the study participant. The comparison of the two data sets should provide indications as to whether anomalies can be measured in the participants. Whether these anomalies are only a product of the individuals or are more generalizable can be determined by comparing the anomalies between the test subjects. As the anomalies are to be detected by an eye tracker, particular focus should be placed on the fixations and saccades of the test subjects as perceived by the eye tracker, as this is the main data that an eye tracker can record.

RQ2: What do users with a need for explanation and time pressure pay particular attention to in order to find an explanation?

The background to this question is that when offering an explanation, it should also be positioned where the user searches for it and finds it. A well-placed explanation helps the user and saves time when searching. Where the user searches probably also depends heavily on the design of the software and the level of experience the user has with software in general. In order to obtain meaningful data for this, it is first necessary to generate a need for explanation among the test subjects. If this need for explanation exists, it could be of interest in the context of RQ2 to see in which areas of interest the eye tracker now measures comparatively more eye movements. In addition, observations by the person conducting the study could indicate where an appropriate place for an explanation would be.

In order to find out whether the eye movements of a technically inexperienced subject behave differently than those of someone with more experience when an explanation is required, you first need a method to assess the level of experience. Once you have determined the level of experience, you can compare all subjects with the same level of experience and compare the average results with the average results of subjects with a higher level of experience. This should show whether a significant difference in eye movement can be determined by different levels of experience.

The idea of this RQ is that there may be a need for personalization to the user when implementing software where eye trackers measure the need for explanation. The answer to RQ3 could indicate an initial factor that should be considered in order to adapt the software to the individual user and make it more functional.

3.2 Methodology

3.2.1 Principle Ideas

To be able to answer the research questions, it is particularly important to provoke a need for explanation from the study participants. At the same time, however, there also had to be parts of the study where the test subjects could complete tasks that did not require any explanation in order to obtain a control data set. The initial study design idea was therefore as follows:

The target duration of the entire study is 30-45 minutes in order not to lose the subject's attention during the course of the study. The tests were prepared in Excel with the help of relatively unknown functions. The reason why I decided to use Excel for this experiment is that I assume that everyone of my participants has used Excel before and knows a few tools of the software superficially, but at the same time this software is so complex that it seems almost impossible to master every function of Excel, which makes it possible to set tasks in the study design that are probably not solvable for a normal user without further help. One test contained functions that many Excel users should know and the other two tests contained very difficult tasks that are difficult to solve without explanation and prior knowledge. After the test person has sat down in front of the computer and signed the data protection declaration, the test person is told that three different tests will be carried out.

In these tests, they have 3 minutes to complete as many tasks as possible in the correct order. Skipping a task is not allowed. It is therefore important to complete the tasks as well as possible and only in second place is it relevant to complete as many of the tasks as possible. Furthermore, it is not permitted to look for explanations on the Internet or to leave the program with performing the test. The study participants may only use the hints that appear when hovering over something with the mouse or the official operating instructions, which will be shown on one half of the screen, if an explanation is required. After each test there is a survey which asks for a few details about the test. There will also be a small demographic survey at the end. While the tests are being carried out, the person conducting the study should also observe whether the study participant has completed the task correctly. If the participant has not yet completed a task but thought it was finished and would like to move on to the next task, the person conducting the study should point this out to the participant and ask them to complete the previous task first. The person conducting the study should not give any explanations to the study participants.

In this way, I prevented the study participants from avoiding the situation where they need explanations by simply completing a later task that is easier for them. It was also deliberately not mentioned that the tasks were to be completed in Excel, in the hope that reduced mental preparation time would increase the need for explanations. The short time of 3 minutes served to avoid putting too much mental strain on the test subjects. The ban on using tools other than those provided was intended to ensure that all test subjects had the same options and do not differ in the explanations used.

In order to give the study participants a feeling for how far they were allowed to move in order to be correctly detected by the eye tracker and to prevent technical errors, the eye tracker was calibrated once at the start of the study so that the participants could see for themselves when the eye tracker could no longer detect them correctly and how freely they were allowed to move. A new calibration was also carried out before each individual test. During the tests, the person conducting the study, who was located behind a second screen opposite the test person, was able to see on his screen where the test person was looking and whether the eye tracker was still able to measure the test person's eyes correctly. If the subject moved their head too far out of the radius of the eye tracker, the person conducting the study pointed this out to the subject and asked them to move back into a better position. The person conducting the study also made notes for each subject if he noticed anything.

3.2.2 Pilot Study and Final Design

To check that the general ideas of the study work, I carried out two pilot studies with employees of the Institute of Software Engineering. It turned out that the study participants paid a lot of attention to the instruction manual on one half of the screen during the difficult tests and managed to reduce their need for explanations. From this I concluded that it is even better if I remove the operating instructions and only allow the explanations that appear when hovering over icons to be used. This ensures that the need for explanation is not simply satisfied and the test subjects continue to search desperately for an explanation.

There was also the idea of mutating the test sequence differently between the study participants, but I intentionally decided against this, because with the expected small number of participants, which is expected in the study, the data would hardly have been comparable and it would probably have been difficult to justify certain results by the respective mutation or the individuality of the test person.

3.2.3 Tests and Surveys

Below you will find the from german to english translated versions of the tests and surveys used for the study.

Test 1

3 Tasks:

1. change the font of the entire text from tile "B5" to "Arial Black"
2. write "Lorem Ipsum" in the "A4" tile
3. give the tile "A4" a font size of 23, make it bold and underlined and

Test 2

4 Tasks:

1. create a new worksheet with the title "Test 1"
2. now automatically calculate the sum of the numbers in tiles "A1" - "D1" from the other worksheet using the sum command in worksheet "Test 1" in tile "B1"
3. create a "stacked bar chart" from the data in the first worksheet in tiles "A5"-"D13" and place the created table in the worksheet "Test 1" in tile "B4"
4. give the chart the title "Consumption of speculoos in parts of Germany by season"

Test 3

4 Tasks:

1. set the name "Costs" for the area "B4"-"B7".
2. in the field "B10", use the sum command to automatically add up the values of all tiles marked with "Costs" without explicitly mentioning the tiles.
3. define a new name exclusively for the worksheet "Table1". The name should be "Object".
4. give the tiles "A4"-"A7" the name "Item"

Survey for All Tests

Select how much you agree with the statements (strongly agree, agree, neutral, disagree, strongly disagree):
- At the beginning of the processing time, I had the feeling that I would be able to complete the tasks in time
- I felt stressed
- I had a high need for explanation

Select the statement that applies most:
- I completed the tasks but could have used some explanation
- I completed the tasks and did not need any explanation
- I did not complete the task because I would have needed an explanation
- I did not complete the task and would only have needed more time (and no explanation)

Demographic Survey

What age group do you belong to?
- 18-21
- 22-25
- 26-28
- 29-32
- 33+

Which gender do you consider yourself to be?
- female
- male
- diverse

Select how much you agree with the statements (strongly agree, agree, neutral, disagree, strongly disagree):
- I consider myself to be very tech-savvy
- I was already familiar with Excel before the study
- I work a lot with computers as part of my work/studies

3.2.4 Execution of the Study

The study was executed without any major problems. The setup of the study environment can be seen in the figure 3.1. The test subject and the person conducting the study sat opposite each other at the same table, each with their own screen. The eye tracker was attached to the screen of the study participant, on which the tests could be seen. On the study coordinator's screen, the participant's eye movements could be seen in real time and the data recording of the eye tracker could also be controlled. During the times when no test was being carried out, the study coordinator went to the study participant's side to answer questions about the questionnaires. In the questionnaires, the study participants were usually not quite sure how to assess the success of their task if they had not completed everything in time 3.2.3. The study coordinator then explained that the most important factor was whether they felt they needed an explanation to complete the test or whether they just needed more time and did not require any guidance.

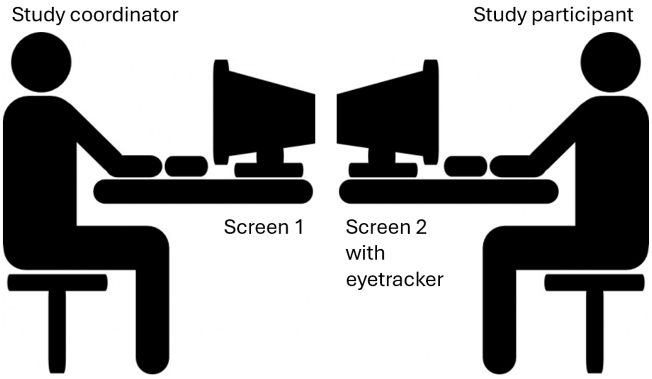
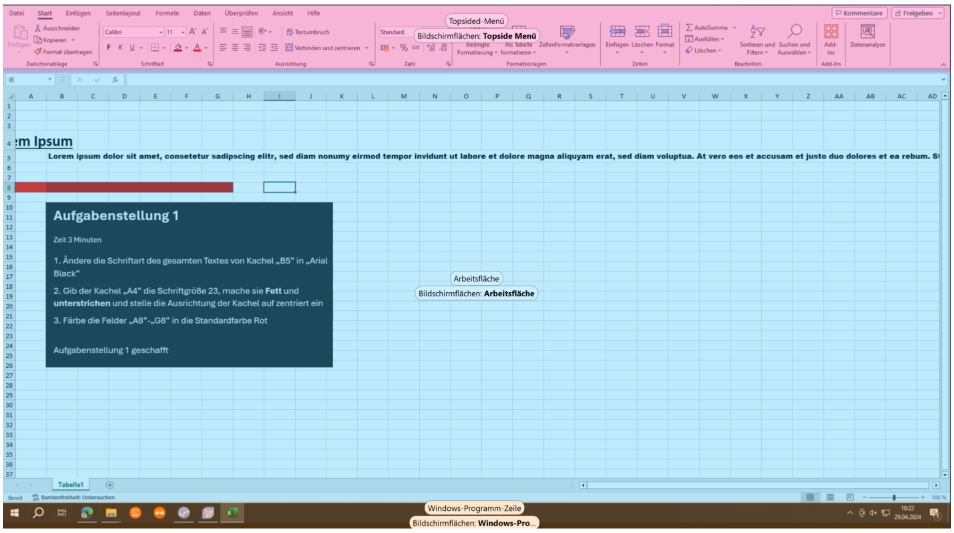


Figure 3.1: Graphical representation of the experimental setup

3.3 Data Analysis Methods

After the study was successfully completed with all eleven subjects, there were 33 screen recordings of the tests with the eye tracking data and the corresponding completed surveys and eleven completed demographic surveys. The Tobii Pro Lab software, which was used to collect the eye tracking data, offers a lot of data that can be exported for analysis. In order to be able to extract the relevant data from the irrelevant data during export, it was necessary to mark the periods of interest in the recordings. These time periods are called Times of Interest (TOIs). The length of these times of interest varies between the tests from less than a minute to 3 minutes, depending on whether and how quickly a test was completed. The resulting data sets have a data point for approximately every 8 ms. Additional information is then output for each data point. I also marked the relevant areas in the recordings, which are called Area of Interest (AOI), where the fixations are assigned if they happened in this area. In the figure 3.2 you can see the AOI's coloured on a screenshot. The yellow area at the bottom is the AOI where the Windows taskbar is located. Above this, in the Excel workbook area, there is another AOI in blue. In pink, you can see the last of the three AOIs at the top, which is used for the Excel toolbar. I have marked these fields as relevant AOIs because they fulfil different purposes and I could not think of any more useful AOIs. During the evaluation, I noticed in 5 of 33 tests that many fixations could not be assigned to any AOIs. After some research and attempts to link the data with the AOIs, I came to the conclusion that it must be a technical error in the software. These tests were not counted in the associated analyses. To answer the research questions, the eye movement type and, in the case of fixations, the AOI are particularly relevant for the evaluation.



em Ipsum

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. St

Aufgabenstellung 1

Zeit 3 Minuten

1. Ändere die Schriftart des gesamten Textes von Kachel „B5" in „Arial Black"

2. Gib der Kachel „A4" die Schriftgröße 23, mache sie **Fett** und **unterstrichen** und stelle die Ausrichtung der Kachel auf zentriert ein

3. Färbe die Felder „A8"-„G8" in die Standardfarbe Rot

Aufgabenstellung 1 geschafft

Figure 3.2: Screenshot with coloured areas of the AOI's

Additional data could have been collected for the evaluation, which would probably have been interesting for answering the research questions. For example, the exact screen coordinates of the fixations or the pupil diameter of the left and right eye would have provided interesting information to answer the research questions. For reasons of simplification and because the tests in the study were not designed for the evaluation of this data, I considered it reasonable to concentrate only on the data mentioned.

As the last section of the evaluation, I correlated and compared the data from the recordings, which were collected with the eye tracker, with the information from the surveys for evaluation 4.4. The resulting findings should provide a good data basis for the discussion 5.

Chapter 4

Evaluation

As part of the evaluation, the answers to the Likert scales were assigned numerical values. "Strongly agree" was given a value of x=1, "agree" a value of x=2, "neutral" a value of x=3, "disagree" a value of x=4, and "strongly disagree" a value of x=5. This assignment enhances comparability between tests and participants and enables the calculation of relevant values, such as the correlation of eye tracker measurements with the respondents' answers.

4.1 Participants Demography

The analyses resulting from the demographic survey are presented in this section.

4.1.1 General Demography

The age and gender of the participants can be seen in Figure 4.1. The sample as a whole was relatively young ($\bar{x} \approx 22.7, \sigma \approx 2.2$). The gender distribution was balanced ($Male \approx 55\%, Female \approx 45\%$).

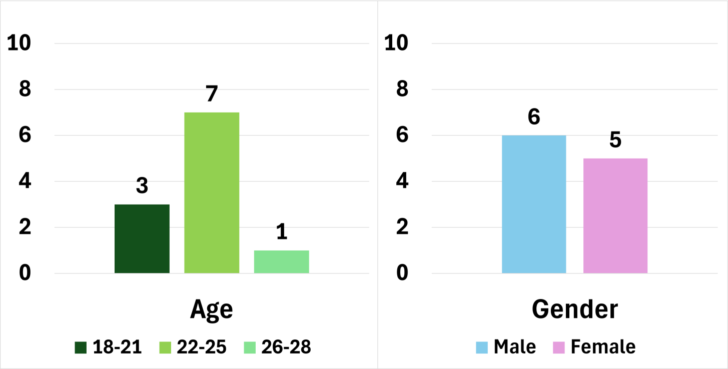


Figure 4.1: **Age and gender** of all 11 participants in the study

4.1.2 Technical Skills of the Participants

Figure 4.2 shows the answers to the study participants' technical self-assessment. The statements they were asked to evaluate are translated from German into English as "I consider myself to be very tech-savvy", "I was already familiar with Excel before the study" and "I work a lot with computers as part of my work/studies". A 5-point Likert scale was used for this.

The majority (55%) see themselves as technically adept, many (36%) were rather neutral on this statement and one person (9%) disagreed. Most (55%) said they had little to no experience in using Excel, some (27%) agreed that they had experience and few (18%) neither agreed nor disagreed. Ten out of eleven participants (91%) work a lot with computers as part of their work/studies, with the exception of one person (9%) who rated themselves as neutral on the statement.

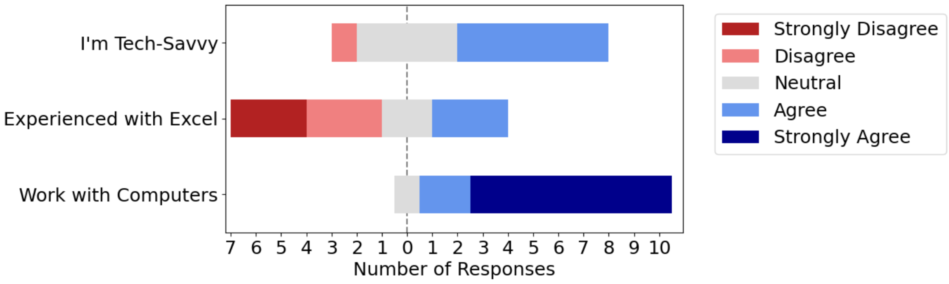


Figure 4.2: Study participants' **technical self-assessment** using a 5-Point Likert scale

Overall, the descriptive statistics (cf. Figure 4.3) of the values from Figure 4.2 show that the greatest difference between the individual test subjects lies in their Excel experience ($\sigma^2 \approx 1.34, \sigma \approx 1.16$) and that the test subjects' self-assessment of their Excel skills is rather bad ($\bar{x} \approx 3.55, \tilde{x} = 4$). In terms of affinity for technology and time spent on the computer, the subjects differed rather little (${\sigma^2}_{\text{Technicel Affinity}} \approx 0.43, \sigma_{\text{Technicel Affinity}} \approx 0.66, {\sigma^2}_{\text{Computer Time}} \approx 0.41, \sigma_{\text{Computer Time}} \approx 0.64$). Also, the test subjects spend a lot of time on the computer as part of their work/studies ($\bar{x} \approx 1.36, \tilde{x} = 1$) and would consider themselves to have a little technical affinity ($\bar{x} \approx 2.55, \tilde{x} = 2$).

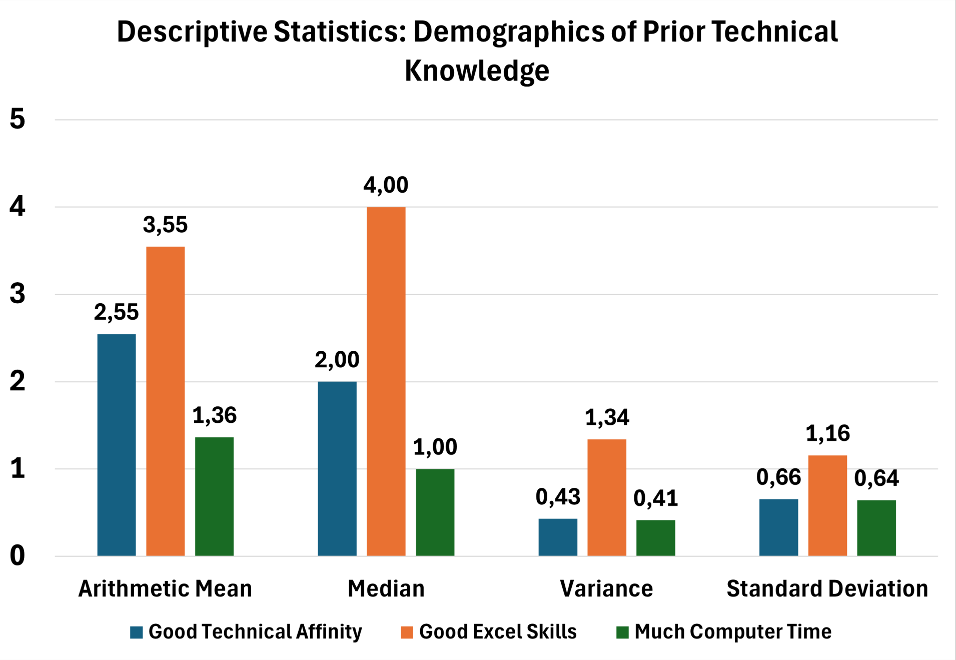


Figure 4.3: Descriptive statistics of the Study participants' **technical self-assessment** using a 5-point Likert scale (cf. Figure 4.2)

4.2 Survey Data After Tests

Here are the analyses of the surveys, which should be completed by the test persons after each test and reflect the test persons' impression of the test. There were 33 tests in total, as each of the three tests was completed by each of the eleven study participants.

4.2.1 Stress Level of Participants During the Study

The respondents' assessments of the statement "I felt stressed" are represented by the figures 4.4 and 4.5.

In the first test, a small proportion of the test subjects (27%) agreed with this statement, while a slightly larger proportion (27% strongly disagreed, 9% disagreed) disagreed with it. The rest (36%) were neutral to the statement.

In the second test, many participants (73%) felt stressed by the test, one person (9%) was neutral and two people (18%) were not stressed by the test.

The last test was relatively balanced with five agreeing, four disagreeing and two neutral answers.

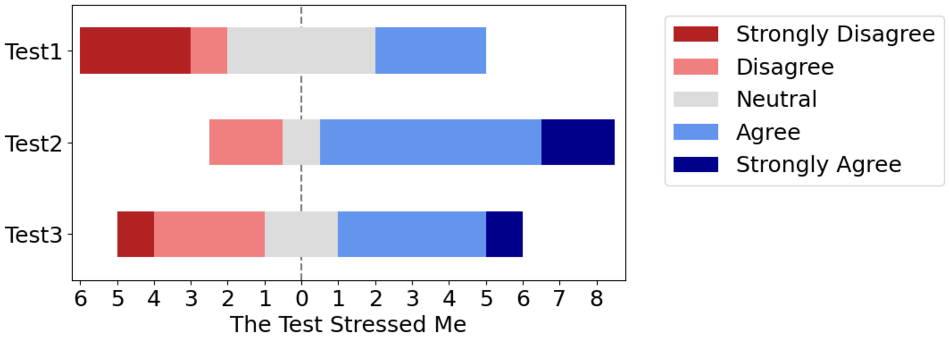


Figure 4.4: Study participants' self-assessment of their **stress levels** during the tests using a 5-point Likert scale

In comparison, Figure 4.5 shows that the first test was the least stressful on average ($\bar{x} \approx 3.36, \tilde{x} = 3$), the third test was slightly more stressful ($\bar{x} \approx 2.91, \tilde{x} = 3$) and the second test was the most stressful overall ($\bar{x} \approx 2.27, \tilde{x} = 2$) for the study participants. The variance and standard deviation differed similarly between test 1 ($\sigma^2 \approx 1.32, \sigma \approx 1.15$), test 2 ($\sigma^2 \approx 0.93, \sigma \approx 0.96$) and test 3 ($\sigma^2 \approx 1.36, \sigma \approx 1.16$).

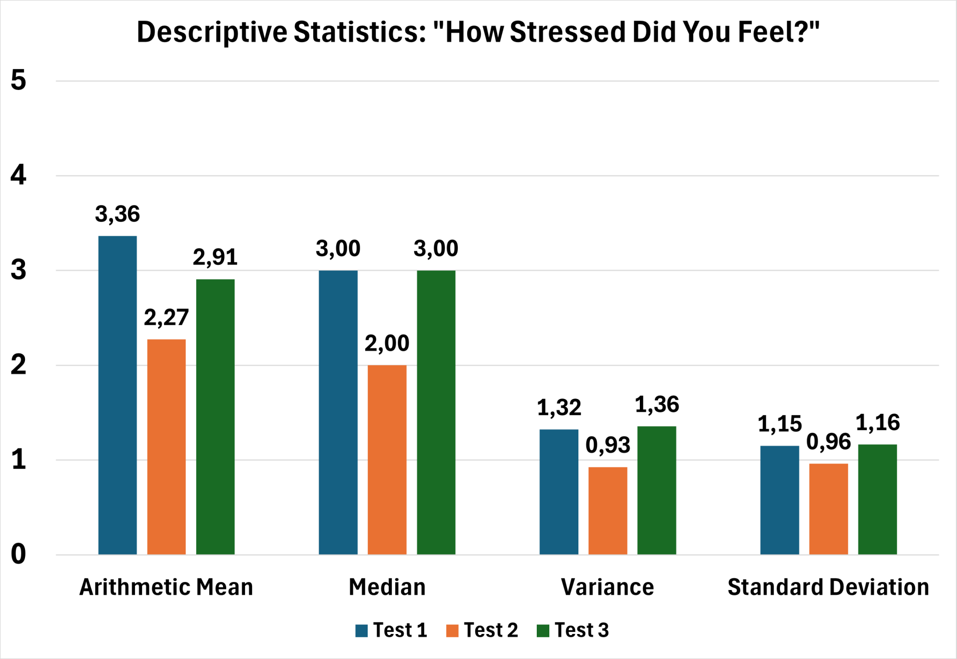


Figure 4.5: Comparison of the descriptive statistics of the study participants' self-assessments of their **stress levels** (cf. Figure 4.4)during the tests, conducted using a 5-point Likert scale

4.2.2 Need for Explanation Level of Participants During the Study

The figures 4.6 and 4.7 represent the participants' answers to the statement "I had a high need for explanation" in the survey.

In 4.6 it becomes clear that the greatest need for explanation was reported by the test subjects in test 2. Seven (64%) people agreed that they needed an explanation. Two further participants (18%) strongly agreed with the statement. Two participants (18%) were neutral to the statement.

In contrast to Test 2, there was the lowest need for explanation in Test 1. Eight people (73%) strongly disagreed with the statement. One person (9%) each responded "disagree", "neutral" and "strongly agree" to the statement.

In the last test - Test 3 - four people (36%) voted "strongly agree" and "disagree". The remaining three study participants (27%) were neutral towards the statement.

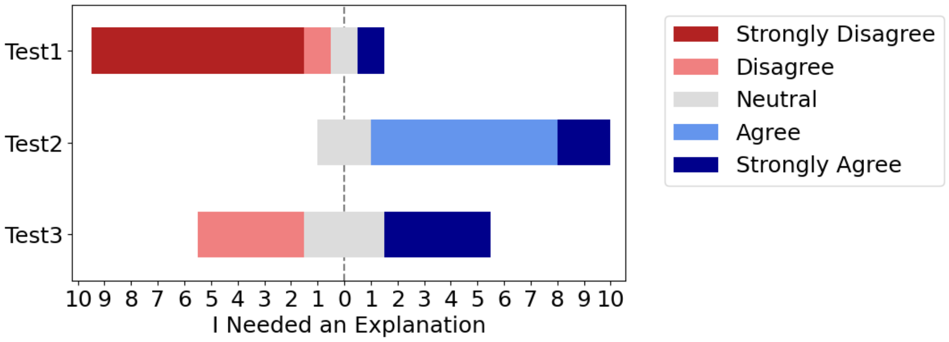


Figure 4.6: Study participants' self-assessment of their **need for explanation** during tests using a 5-point Likert scale

The descriptive statistics (ct. Figure 4.7) to Figure 4.6 show that, all in all, the first test required the least explanation ($\bar{x} \approx 4.36, \tilde{x} = 5$), the second test required the most explanation ($\bar{x} \approx 2.64, \tilde{x} = 3$) and the third test required slightly less explanation than the second test ($\bar{x} = 2, \tilde{x} = 2$). The variance and standard deviation were very low in the second test($\sigma^2 \approx 0.36, \sigma \approx 0.6$), while they were comparatively high in the first and third tests(${\sigma^2}_{\text{Test1}} \approx 1.5, \sigma_{\text{Test1}} \approx 1.23, {\sigma^2}_{\text{Test3}} \approx 1.69, \sigma_{\text{Test3}} \approx 1.3$).

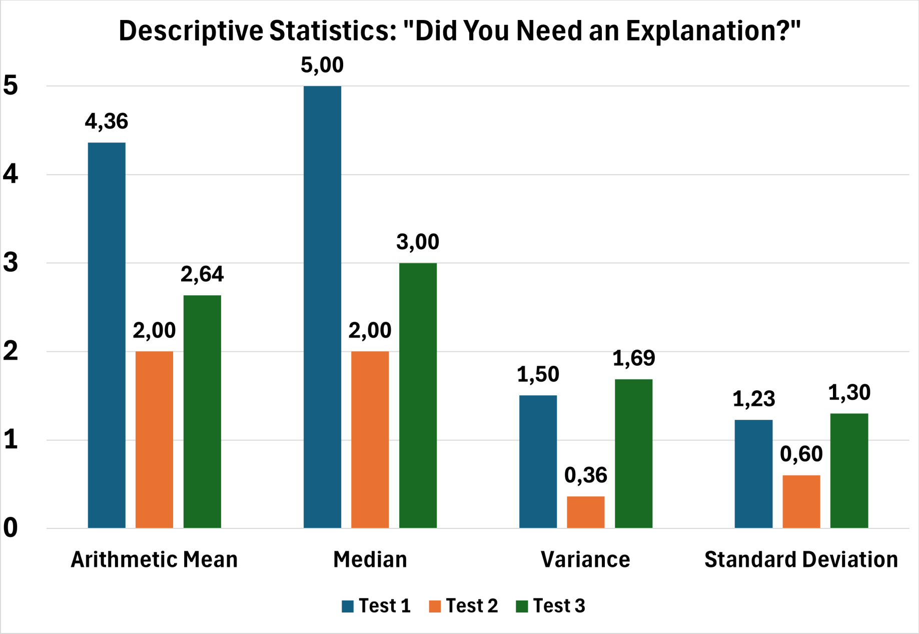


Figure 4.7: Comparison of the descriptive statistics of the study participants' self-assessments of their **need for explanation** (cf. Figure 4.6)during the tests, conducted using a 5-point Likert scale

4.2.3 Feasibility of Tests Estimated by the Test Subjects During the Study

The figures 4.8 and 4.9 reveal the results of the Likert survey on the feasibility of the tests.

The first test was considered feasible by 9 people (64% strongly agreed, 18% agreed). One person (9%) was neutral and two people (18%) disagreed.

In the second test, five people (27% strong disagreement, 18% disagreement) found the test very difficult to solve in time, four people (27% agreement, 9% strong agreement) found the test feasible in time and two people (18% neutral) could not be categorized in either direction.

Most of the participants' responses to the third test disagreed with the statement. Three participants (27%) each "agreed", were "neutral" or "disagreed". Two participants (18%) "strongly disagreed".

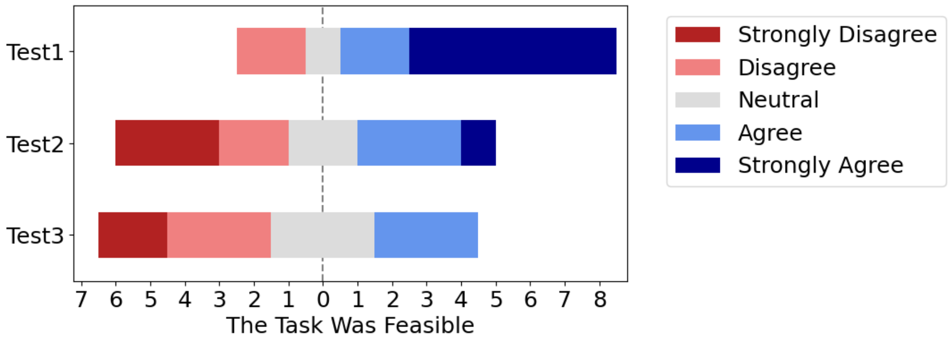


Figure 4.8: Study participants' assessment of the **feasibility** of the tests using a 5-point Likert scale

To summarize (cf. Figure 4.9), the first test was considered very manageable ($\bar{x} \approx 1.91, \tilde{x} = 1, \sigma^2 \approx 1.36, \sigma \approx 1.16$). The opinions on the second and third test were relatively similar and categorized the tests as rather not feasible ($\bar{x}_{\text{Test2}} \approx 3.27, \tilde{x}_{\text{Test2}} = 3, \sigma^2{}_{\text{Test2}} \approx 1.83, \sigma_{\text{Test2}} \approx 1.35, \bar{x}_{\text{Test3}} \approx 3.36, \tilde{x}_{\text{Test3}} = 3, \sigma^2{}_{\text{Test3}} \approx 1.14, \sigma_{\text{Test3}} \approx 1.07,$).

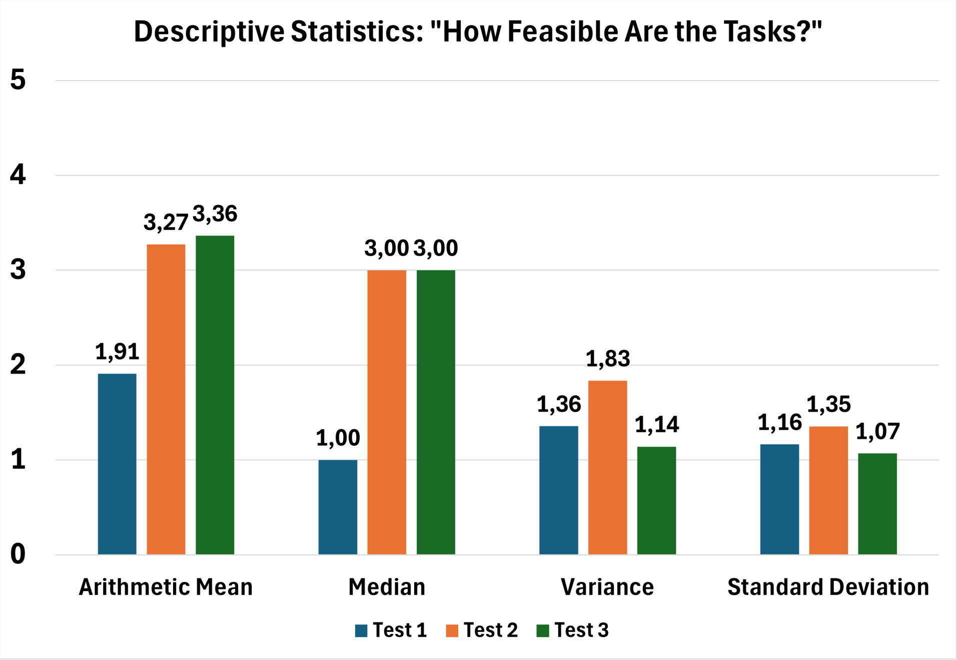


Figure 4.9: Comparison of the descriptive statistics of the study participants' assessment of the **feasibility** (cf. Figure 4.8) of the tests, conducted using a 5-point Likert scale

4.2.4 Test Subjects' Statements on the Success of Individual Tests

As part of the study, after completing a test, the participants were asked to tick one of four pre-formulated sentences in the survey that best reflected their personal results. If none of the four predefined sentences fitted, there was also the option of ticking a fifth box and formulating a more suitable sentence to the study coordinator. As this option was not used once by a respondent, it is not considered in the analysis or in Figure 4.10.

The statements in the legend of Figure 4.10 are short for the following sentences: "Didn't Need Any Explaination" stands for "I completed the tasks and did not need any explanation". "Could Need Explaination" stands for "I completed the tasks but could have used some explanation". "Need for More Time" stands for "I did not complete the task and would only have needed more time (and no explanation)". "Need for Explanation" stands for "I did not complete the task because I would have needed an explanation".

As can be seen from Figure 4.10, ten people (91%) completed the tasks from test 1 in full. One of these ten (9% from all eleven) could have used an explanation to solve the tasks. The one person (9%) who did not complete the tasks in the specified time stated that they only needed more time and no explanation.

In the second test, one person (9%) stated that they had completed all the tasks in the given time, although they could have used an explanation. Of the ten people (91%) who did not manage to solve all the tasks in the given time, four people (36% of all eleven) stated that they would have needed more time and six participants (55% of all eleven) would have needed an explanation to solve the task.

As with the second, only one participant (9%) stated that they had completed the third task but could have used an explanation. Of the other ten study participants (91%) who did not complete the tasks in time, six participants (55% of all eleven) stated that they only needed more time, while the other four participants (36% of all eleven) stated that they needed an explanation to solve the tasks.

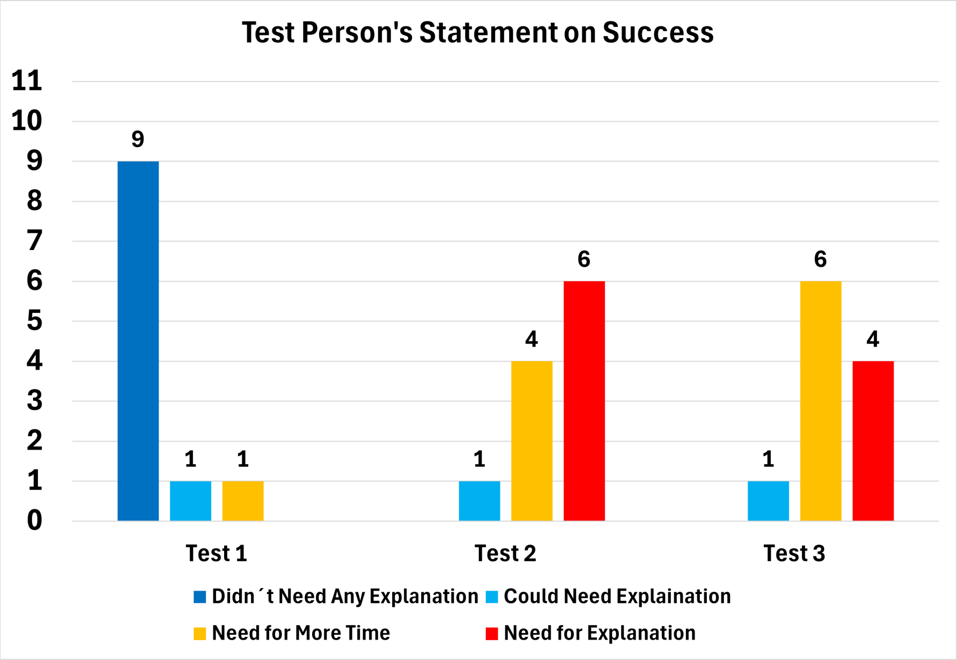


Figure 4.10: Statements from the study participants on **how they completed the individual tests**

4.3 Data Collected Through Eye Tracker During the Study

The figures 4.11 and 4.12 shown in this section is intended to provide an overview of whether there are significant abnormalities between the tests as a whole when comparing all eye tracker data for the three individual tests.

For fixations, the percentage of all eye movements was highest for Test1 (67%) and Test3 (67%), closely followed by Test2 (64%). If you look at the results with the second decimal place, Test1 (67.35%) has a slightly higher percentage compared to Test3 (66.9%). In the saccades, it can be seen directly that Test 1 (17%) has a higher proportion of all eye movements compared to Test 2 and 3 (15%). If you round the saccades to the second decimal place, you can see that Test 3 (15.11%) has a slightly higher proportion of saccades compared to Test 2 (14.95%). The data points in the data set where either the gaze could not be captured correctly by the eye tracker or could not be categorised correctly were summarised under "Other". Here it is clear that Test 1 (16%) has the smallest, Test 2 (21%) the largest and Test 3 (18%) an overall proportion between the other two tests.

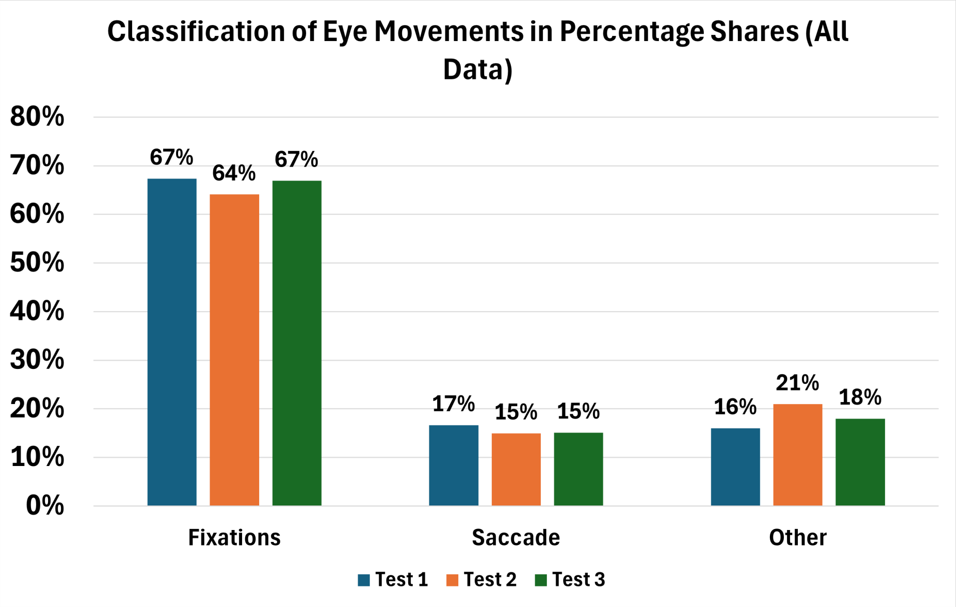


Figure 4.11: For each individual test, the **cumulative classified eye movement data** of **all 11 study participants**, presented as a **percentage of the total** data for each test

If you only look at the proportion of fixations compared to saccades and exclude the "Others" from the comparison, you get the data shown in Figure 4.12. Here it is clear that the proportions of fixations to saccades only differ by a small proportion (between 1-2%) between the tests. Test 1 has the lowest proportion of fixations (80%) and Test 3 (82%) the highest.

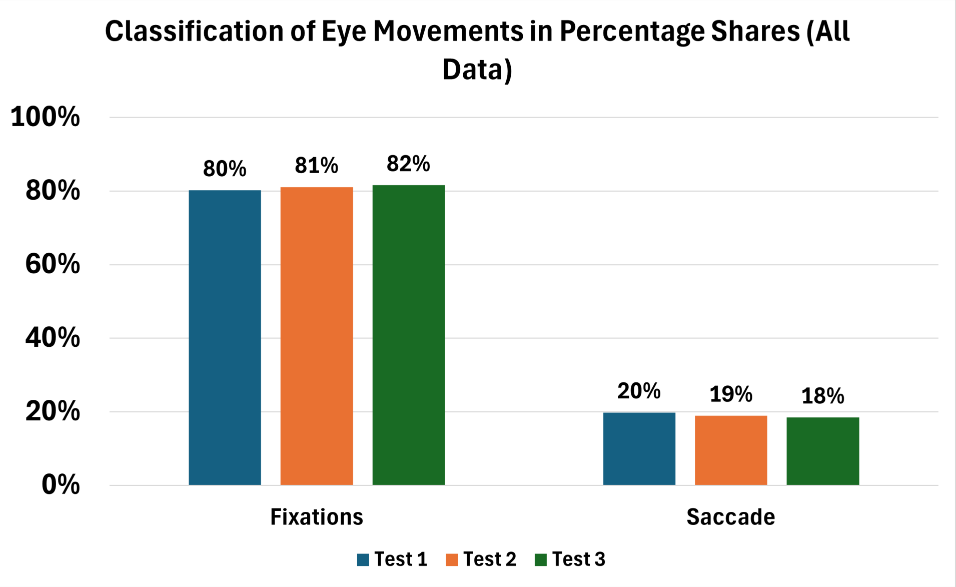


Figure 4.12: For each individual test, the **cumulative classified eye movement data** of **all 11 study participants**, presented as a **percentage of the total** data for each test

4.4 Survey and Eye Tracker Data Correlation

In this section of the analysis, the data from the previous sections are correlated. In all figures in this section, the numbers from 1-5 on the x-axis stand for 1="strongly agree" to 5="strongly disagree", as explained at the beginning of the chapter 4.

The figures 4.14, 4.16, 4.18, 4.20, 4.22, 4.24 and 4.25 are rather unusual. The correlations in these always refer to the three tests that each participant fulfilled individually. This means that each bar represents the correlation of the data of a single respondent with the value of a Likert survey. The Correlation coefficient can be seen on the bar. The significance is stated in the associated text. Accordingly, the degree of freedom for all these tests is always 1.

4.4.1 Correlation Between Explainability and Eye Movement

The Figure 4.13 shows the percentage of fixations in correlation with the subjects' indication of how high their need for explanation was when completing the test. The two variables do not appear to correlate ($r(31) = .09, p > .05$).

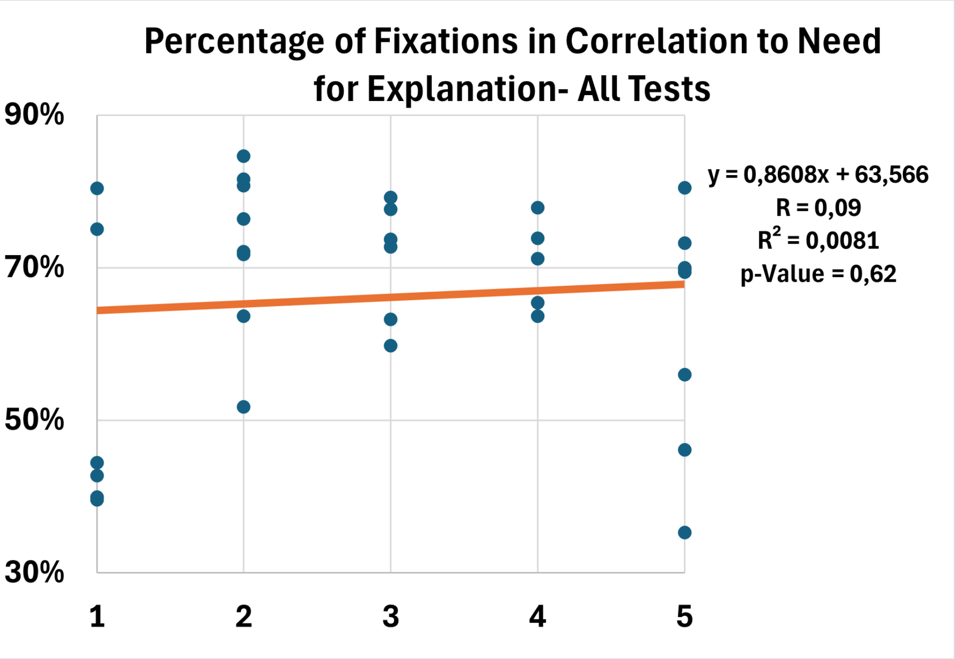


Figure 4.13: Change in the proportion of **fixations** for **all tests** correlated with the level of **need for explanation** reported by each study participant after completing the test

In Figure 4.14, the data from Figure 4.13 are calculated again individually for each participant. The results from participants 1, 4, 5, 8, 9, 10 and 11 (64%) have an extremely significance ($p < .001$). The correlation of participants 6 and 7 (18%) also have a high significance ($p < .01$). In contrast, the correlation coefficients of participants 2 and 3 (18%) have no significance ($p > .05$).

Figure 4.15 illustrates the relationship between the percentage of saccades and the subjects' reported need for explanation while completing the test. The correlation between the two variables is not significant ($r(31) = .0735, p > .05$).

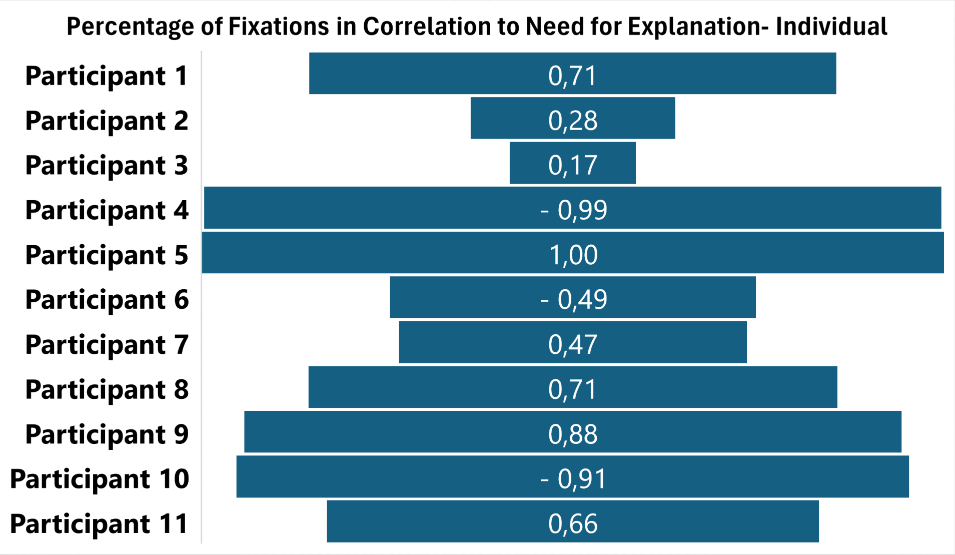


Figure 4.14: Change in the individual proportion of **fixations** across the **three tests performed by the participant**, correlated with the level of **need for explanation** reported by the study participant after completing each test

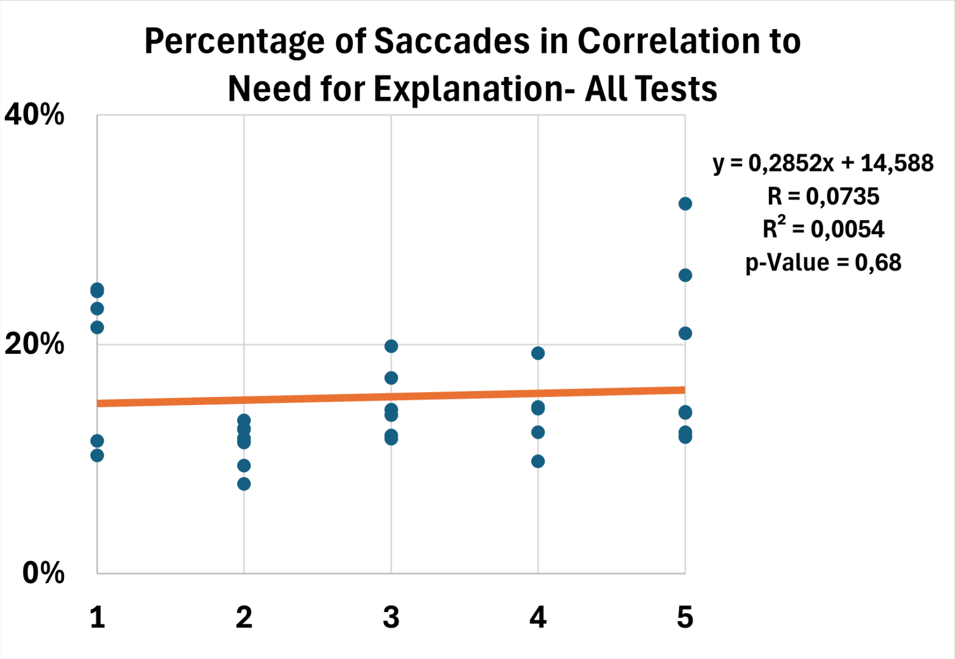


Figure 4.15: Change in the proportion of **saccades** for **all tests** correlated with the level of **need for explanation** reported by each study participant after completing the test

In Figure 4.16, the data from Figure 4.15 are recalculated for each individual participant. The results for participants 1, 2, 3, 4, 5, 6, 9 and 10 (73%) show extremely significant correlations ($p < .001$). The correlation for participant 7 (9%) is also highly significant ($p < .01$). Conversely, the correlation coefficients for participants 8 and 11 (18%) are not significant ($p > .05$).

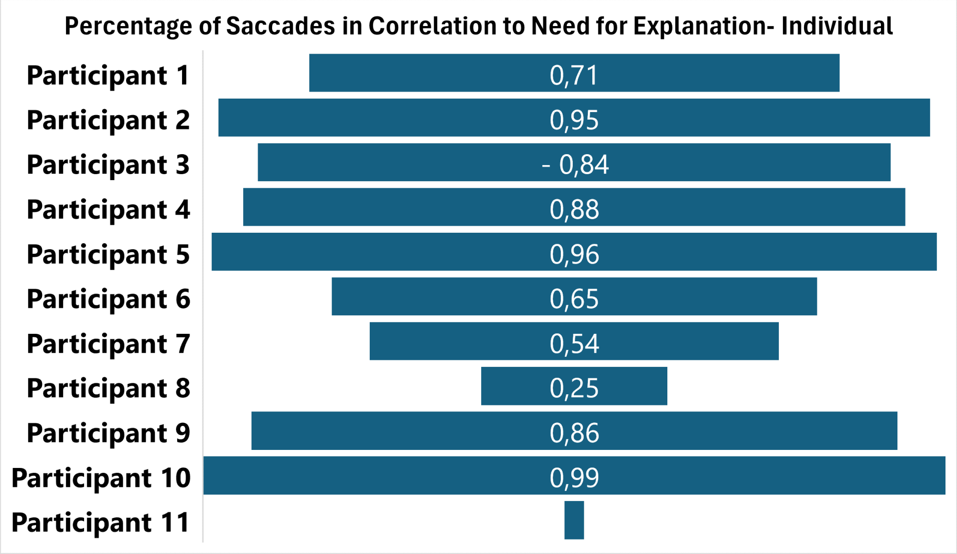


Figure 4.16: Change in the individual proportion of **saccades** across the **three tests performed by the participant**, correlated with the level of **need for explanation** reported by the study participant after completing each test

4.4.2 Correlation Between Stress and Eye Movement

Figure 4.17 displays the percentage of fixations in relation to the subjects' self-assessed stress during the test. The analysis shows no significant correlation between the variables ($r(31) = .3121, p > .05$).

Figure 4.18 depicts the individually for every participant recalculated data from Figure 4.17. Participants 3, 4, 6 and 8 (36%) show extremely significant results ($p < .001$). Participants 7 and 9 (18%) demonstrate highly significant correlations ($p < .01$). The participant 10 (9%) show a slightly significant correlation to stress ($p < .05$). However, participants 1 and 11 (18%) do not show significant correlations ($p > .05$). Participants 2 and 5 (18%) did not show any correlation, as they reported having the same stress level across all three tests. This is also the case for Figure 4.20 and Figure 4.22.

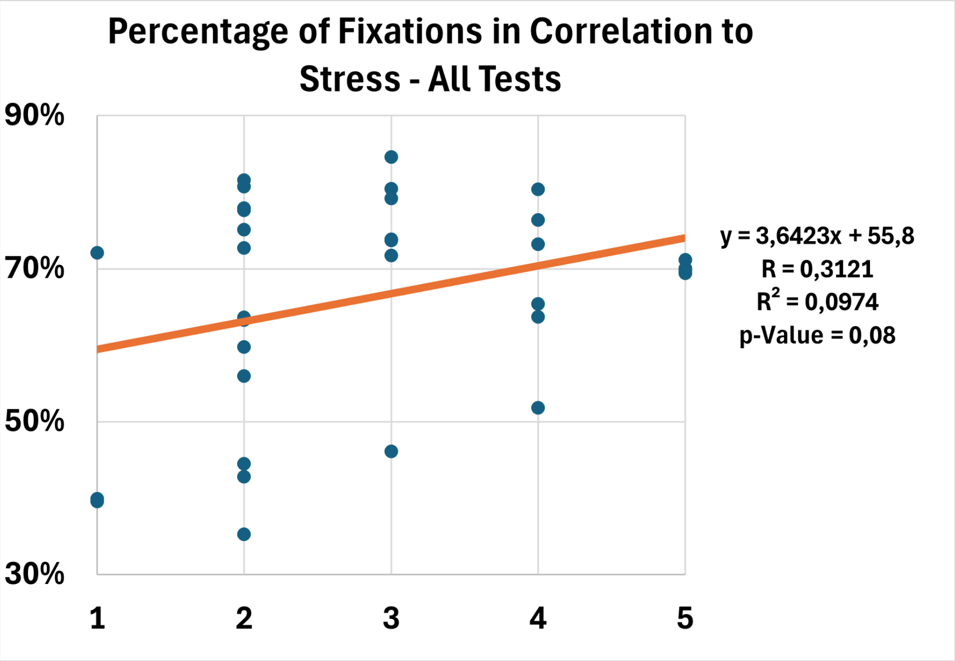


Figure 4.17: Change in the proportion of **fixations** for **all tests** correlated with the level of **stress** reported by each study participant after completing the test

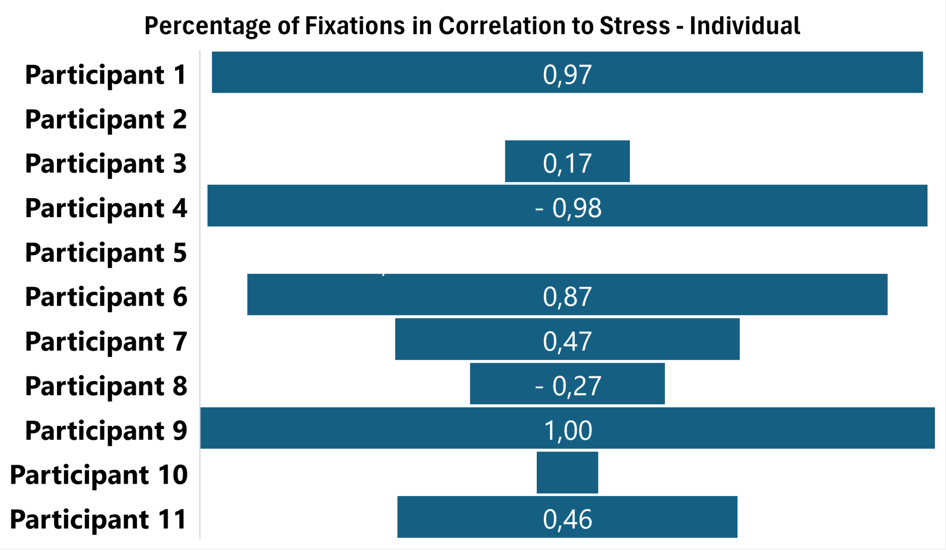


Figure 4.18: Change in the individual proportion of **fixations** across the **three tests performed by the participant**, correlated with the level of **stress** reported by the study participant after completing each test

In Figure 4.19, the correlation between the percentage of saccades and the subjects' perceived stress level when taking the test is shown. The results indicate a significant relationship between the two variables ($r(31) = -.4032, p < .05$).

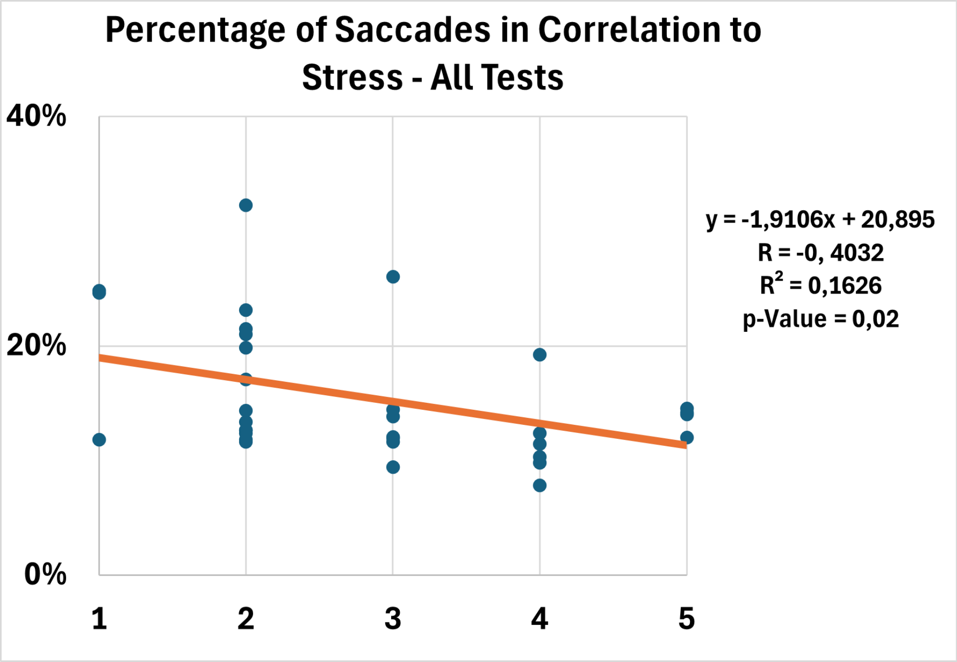


Figure 4.19: Change in the proportion of **saccades** for **all tests** correlated with the level of **stress** reported by each study participant after completing the test

As shown in Figure 4.20, the data from Figure 4.19 have been recalculated on an individual basis. Participants 1, 4, 6, 7, 8, 9 and 10 (64%) exhibit extremely significant correlations ($p < .001$). Participants 3 and 11 (18%) also have a significant correlation ($p < .05$).

Figure 4.21 presents the correlation between the percentage of invalid data (Unclassifiable or measurable eye movement) and the subjects' indicated stress level when completing the test. The analysis reveals no significant relationship between the two variables ($r(31) = -.2127, p > .5$).

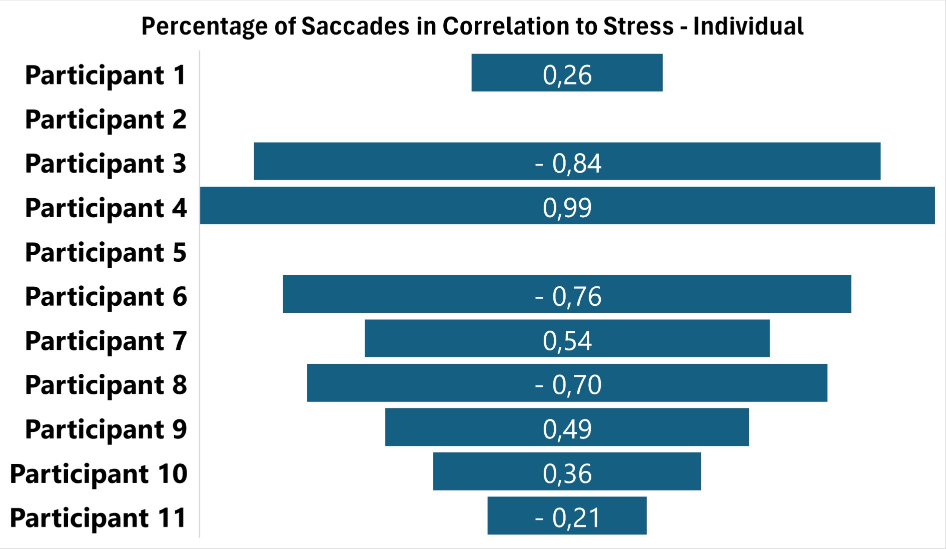


Figure 4.20: Change in the individual proportion of **saccades** across the **three tests performed by the participant**, correlated with the level of **stress** reported by the study participant after completing each test

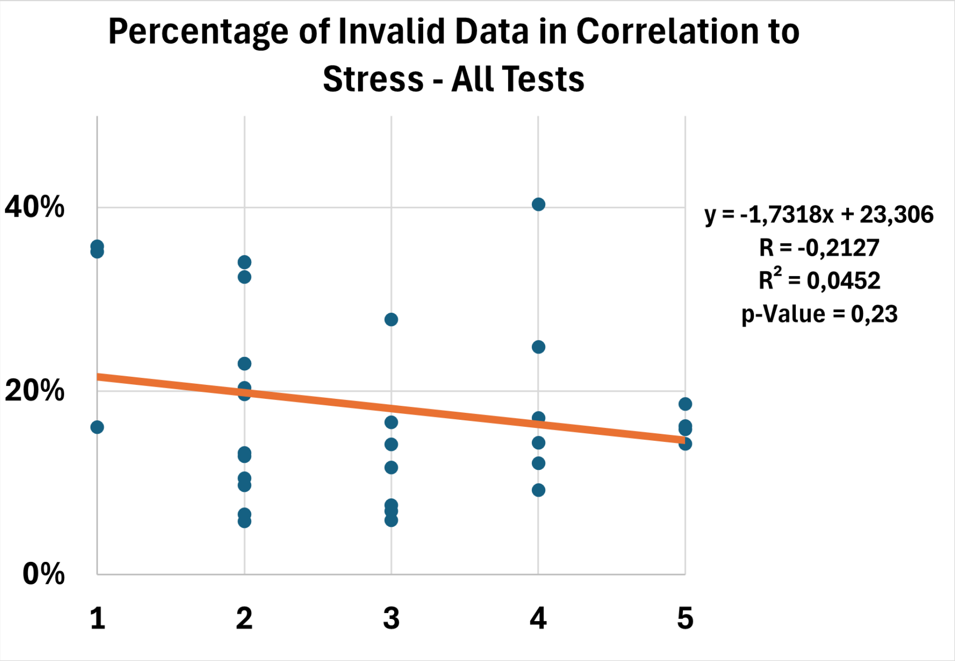


Figure 4.21: Change in the proportion of **invalid data** for **all tests** correlated with the level of **stress** reported by each study participant after completing the test

In Figure 4.21, individual recalculations of the data from Figure 4.22 are presented. Extremely significant results are observed for participants 1, 4, 6 and 9 (36%) ($p < .001$). Participants 7 and 11 (18%) show high significance ($p < .01$). Participants 3, 8 and 10 (27%) have non-significant correlation coefficients ($p > .05$).

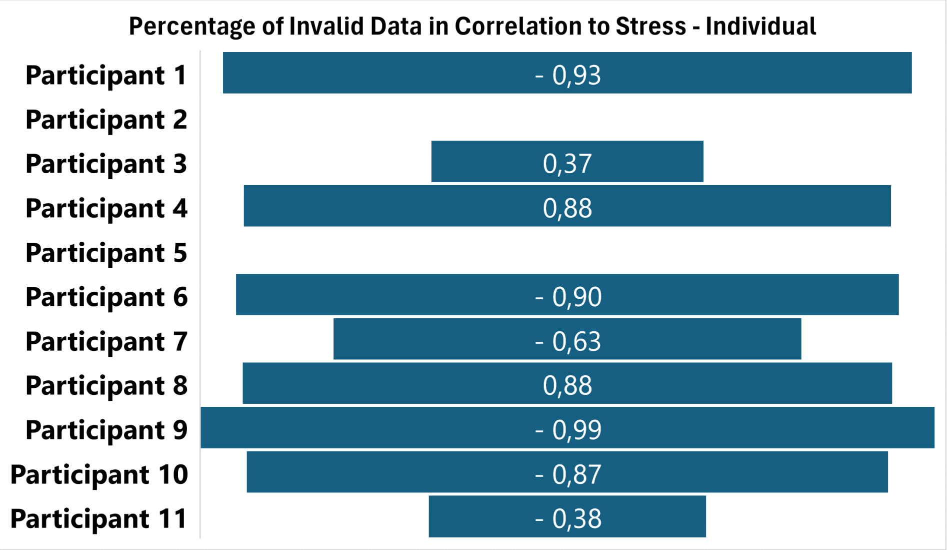


Figure 4.22: Change in the individual proportion of **invalid data** across the **three tests performed by the participant**, correlated with the level of **stress** reported by the study participant after completing each test

4.4.3 Correlation Between Explainability and Areas of Interest

The percentage shares of Fixations in the different AOI regions and the subjects' stated need for explanation during the test are plotted in Figure 4.23. This diagram excludes all tests where more than 10% of all fixations could not be assigned to any AOI. This affected a total of 5 test results, which were removed from the evaluation. There appears to be no significant correlation between the need for explanation and saccades located on the windows taskbar ($r(26) = -.2017, p > .05$). The few fixations that were counted for both AOI's Workbook and Menu bar also show no significant correlation in relation to the need for explanation ($r(26) = -.1427, p > .05$). There is a very significant correlation between the percentage of fixations in the workbook area and the need for explanation ($r(26) = -.5525, p < .01$). The percentage of fixations on the menu bar also has a significant correlation with an increasing need for explanation ($r(26) = .5477, p < .01$).

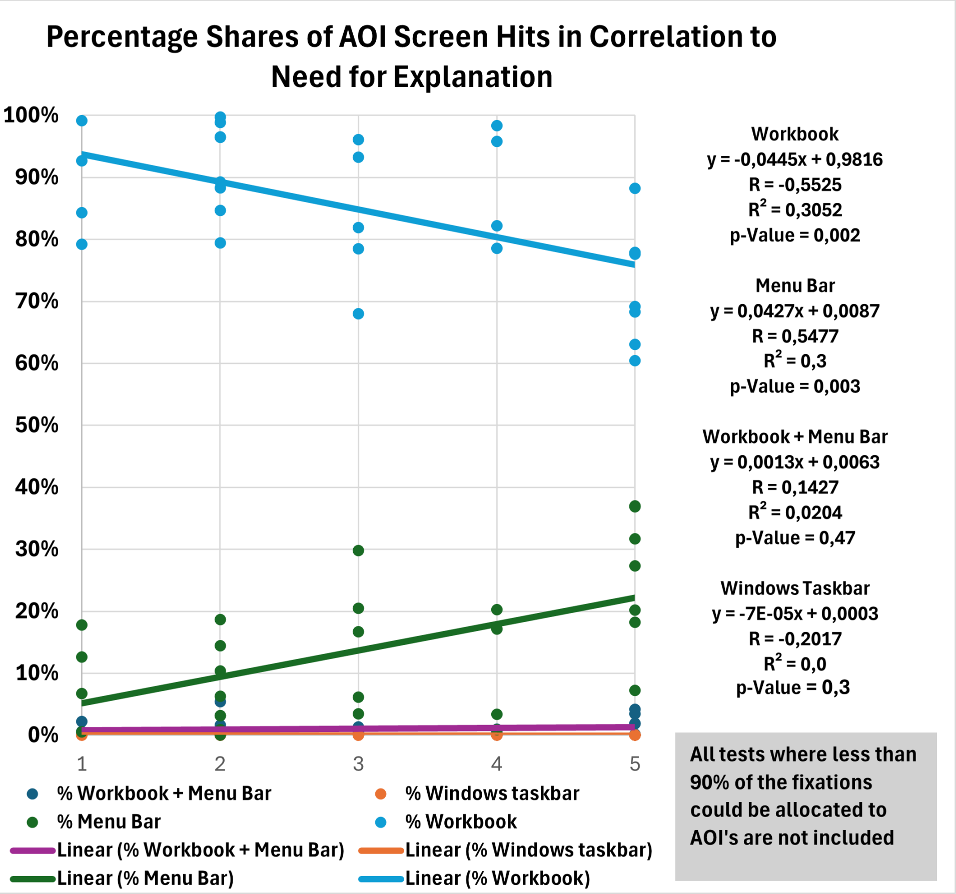


Figure 4.23: Change in the proportion of **all Fixations in certain AOI'S** for **all tests** correlated with the level of **need for explanation** reported by each study participant after completing the test

Figure 4.24 presents data recalculated individually for each participant based on the AOI "menu bar" data from Figure 4.23. All participants except participant 2 (86% of all seven) exhibit extremely significant results ($p <$ .001). Participant 2 (14% of all seven) do not display significant correlations ($p >$ .05).

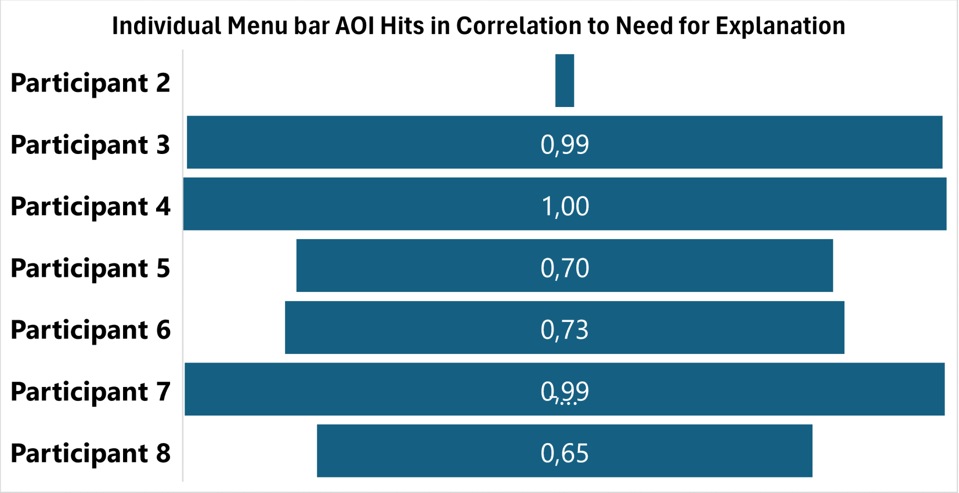


Figure 4.24: Change in the individual proportion of **fixations in the AOI "Menu bar"** across the **three tests performed by the participant**, correlated with the level of **need for explanation** reported by the study participant after completing each test

As illustrated in Figure 4.25, the AOI "workbook" data from Figure 4.23 are individually recalculated for each participant. As in Figure 4.24, all correlations of the test subjects have an extremely high significance with one exception for test subject 2 (86% of all seven) ($p < .001$). Subject 2 (14% of all seven) has no significant correlation ($p > .05$).

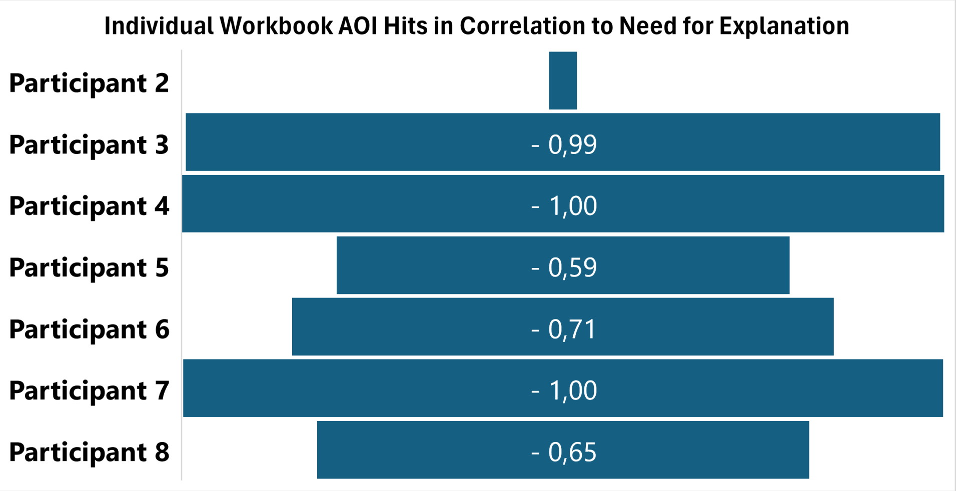


Figure 4.25: Change in the individual proportion of **fixations in the AOI "Workbook"** across the **three tests performed by the participant**, correlated with the level of **need for explanation** reported by the study participant after completing each test

4.5 Observations During the Study

There were a few observations during the study that seem worth mentioning. The person conducting the study noticed that the test subjects who were at a loss often used two specific methods. The people who seemed more experienced used Excel's context menu to find a way to solve the tasks. Others read through the task more often. Only a few searched the menu bar in search of a solution.

Chapter 5

Discussion

5.1 Answering the Research Questions

RQ1: What data provided by an eye tracker can be used to detect irregularities in the behaviour of test subjects when explanations are required?

What quickly becomes apparent when answering RQ1 is that there is no generally significant correlation between the subjects' personal need for explanation and their eye movements. The individual need for explanation, as well as the resulting eye movements, differ noticeably between the individuals of the test subjects. This becomes clear when you look at the results shown in Figures 4.13 and 4.15, which indicate that there is no correlation between the estimated need for explanation and the change in the proportion of saccades and fixations. Although Figures 4.7 and 4.10 show quite well that tests 2 and 3 triggered significantly more need for explanation in the test subjects than test 1, if one looks at the proportions of saccades and fixations of all tests together (cf. Figure 4.12), only very small changes between the proportions of the eye movement classes can be perceived, which also do not change uniformly with the need for explanation.

Interestingly, however, a correlation of fixations and saccades with a need for explanation can be found individually in a large proportion of the test subjects. Figures 4.14 and 4.16 show a very strong correlation with an extraordinarily high significance, which indicates that it may be possible to determine a subject's need for explanation relatively accurately on the basis of a few preliminary measurements.

With the assumption, which is also supported by Deters et al. [13], that the need for explanation also influences the stress level, it is also interesting to see how much the eye movements correlate with the level of stress. The correlation of fixations with stress (cf. Figure 4.17) has a rather weak significance (p<.1), while the correlation of saccades with stress (cf. Figure 4.19) shows a high significance (p<.05). In the individual calculations of the correlation of stress to saccades and fixations, significant dependencies can also be recognized. It also appears that the amount of invalid data collected by the eye tracker is increased by stress (cf. Figures 4.11, 4.21, 4.22). However, it would be wrong to say that the stress is caused by the need for explanation, as it may also have been influenced by the study environment and other factors that were not actively observed as part of the study. In order to make a concrete statement here, more detailed investigations would be necessary, but this gives a further indication of how the need for explanation could be determined for the user.

A last interesting factor that was observed is that there is a significant correlation between the need for explanation and the frequency with which a particular field in the software is fixated with the eyes. In Figures 4.23, 4.24 and 4.25 it can be seen very clearly that all test subjects had more fixations in the work area of Excel and fewer fixations in the menu bar of Excel when there was an increased need for explanation. This could also have been due to the fact that the task description was displayed in the Excel workspace and users stared at the task description more often when they were stuck. This suggests that it should be possible to use an eye tracker to determine when the user needs an explanation when using specific kinds of software.

In summary, the need for explanation seems to be best measured by the individual amount of saccades and fixations in the test person. However, the proportions of saccades to fixations seem to change differently for each subject depending on the need for explanation, which means that it would have to be specified for each user. A further possibility for identifying the need for explanation is offered by the AOI's which are visited particularly regularly by the test person when using the software if there is a need for explanation. The results also show that stress correlated better with the data from an eye tracker than the need for explanation. If the causality of the need for explanation and stress is investigated in more detail, stress could also prove to be an indicator of the need for explanation that can be measured by the eye tracker.

RQ2: What do users with a need for explanation and time pressure pay particular attention to in order to find an explanation?

RQ2 is specifically answered for the Excel software in this study. As already mentioned in the context of RQ1, Figures 4.23, 4.24 and 4.25 show that with an increased need for explanation in Excel, most test subjects focused more on the workbook and less on the menu bar. However, this was probably also due to the fact that the text tasks to be completed as part of the tests were placed in text form in the workbook. Some of the test subjects then read through the task several times when they were stuck, presumably to notice a hint indicating how they could complete the task.

Another reason why most of the test subjects looked more often at the workbook and not at the menu bar when they needed an explanation is that some of them opened the context menu by right-clicking on a tile when searching for a suitable function to solve the tasks and read every function from the top down until they found one that seemed useful. The fixations recorded in the process should therefore all have been credited to the AOI of the workbook. It is also important to note here that the method with the context menu was mainly used by more technically experienced test subjects, who also seemed to have a little more prior knowledge of Excel.

As a result, the answer to RQ2 in the context of this study is the text of the task and the context menu of Excel. More generally, if you look at more than just Excel, a task definition, if it exists, is probably still a very popular fixation point. It is also reasonable to assume that the context menu is often used by many people, especially those with technical experience, to find help while using the software.

5.2 Limitations and Threats to Validity

Some factors may limit the validity of the results of this study. In particular, the participants and the study design must be considered.

The study was conducted with a total of 11 German test subjects who were no older than 28. The low number of participants means that the results are not representative of the population as a whole. However, the aim was not to obtain a result for the entire population, but primarily to give an insight into what is possible when using an eye tracker. In the scope of this work, it would not have been possible to include more test subjects. The fact that all participants are German could also have a cultural influence on the approach to the tests and the eye movements.

The study design of the work was strongly influenced by the Excel software. It cannot be ruled out that the results would have been different if other software had been used for the study. A bias due to the order of the tests can also be assumed. The reason why the order of the tests was not randomised from subject to subject was due to the small number of participants, which would have made it impossible to determine whether the result had changed due to the altered test sequence or due to the individuality of each subject.

When analysing the results, the correlations calculated individually for each subject should be viewed with caution, as these calculations were carried out on a sample size of 3 ($df = 1$).

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In order to find out whether explainability needs can be detected by the means of an eye tracker, this thesis developed a study to trigger the need for explanation in test subjects. During the study, the test subjects had to solve 3 tests in Excel under time pressure and were not allowed to use any external help to solve the tasks. After each test, the test subjects had to complete a survey in which, among other things, their stress levels and need for explanation during the test was assessed. The measured data gave some indication that the need for explanation can indeed be determined with the help of an eye tracker. The results showed that there were no significant eye movements due to the need for explanation that applied to every test subject. However, it was found that individual behaviours were recognisable in most of the test subjects, which made it possible to identify a connection between the eye movements and the subjects' need for explanation. It was also found that stress, which is often used as an indicator for the need for explanation, is particularly easy to recognise with an eye tracker. It was also possible to make statements about where subjects are more likely to look when they need explanations.

All in all, this work shows that the need for explanation can be recognised in different ways with an eye tracker and that the need for explanation has an individual effect on the eye movements of test subjects.

6.2 Future Work

There are many points that could require closer examination in the context of this work in order to be more meaningful. I consider the most important points to be the following:

Recognise the Need for Explanation for a Single Individual as Precisely as Possible

As was noticed during the evaluation, the correlation between eye movements and the need for explanation is especially noticeable in individual participants. In order to evaluate this in more detail, it would be useful to conduct a study with the aim of precisely determining the strength of the subject's need for explanation based on the data that an eye tracker can measure.

This would first require a few tests to recognise how the subject's eye movements change depending on the need for explanation. You could then try to carry out further tests with the same subject and estimate the need for explanation before comparing the estimate with the need for explanation stated by the subject.

Differentiate Between Stress and the Need for Explanation

As it turned out, it is quite possible to recognise stress in the test person with an eye tracker. As stress can be caused not only by the need for explanation, but also by other factors, it could be useful to try to separate the two when using an eye tracker.

Creating another study comparing how subjects' eye movements behave when they are only stressed, stressed by a need for explanation or when they only have a need for explanation could be informative about how well you can separate the two, or whether this is impossible with the data provided by an eye tracker.

Software that Provides Explanations at the Right Moment

An interesting idea to see if the previous proposition is possible would be to build software that provides an explanation the moment the eye tracker recognises the need for an explanation. A prototype that is kept very minimalist and does not require much explanation would probably be useful to start with. Expanding this prototype once you realise that basic concepts work, so that it also recognises what explanation is needed and doesn't just spit out any explanation would be an exciting project that could raise the explainability of software to a new level.

Bibliography

[1] Tobii Connect eye tracker calibration and validation. `https://connect.tobii.com/s/article/eye-tracker-calibration?language=en_US`. Published on: 2023-Jun-26, Accessed on: 2024-June-04.

[2] Tobii Connect how do tobii eye trackers work. `https://connect.tobii.com/s/article/How-do-Tobii-eye-trackers-work?language=en_US`. Published on: 2024-Jan-09, Accessed on: 2024-June-04.

[3] P. Baxter, S. Jack, et al. Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4):544–559, 2008.

[4] J. Bloomfield and M. J. Fisher. Quantitative research design. *Journal of the Australasian Rehabilitation Nurses Association*, 22(2):27–30, 2019.

[5] W. Brunotte, A. Specht, L. Chazette, and K. Schneider. Privacy explanations – a means to end-user trust. *Journal of Systems and Software*, 195:111545, 2023.

[6] L. Carvajal, A. M. Moreno, M.-I. Sánchez-Segura, and A. Seffah. Usability through software design. *IEEE Transactions on Software Engineering*, 39(11):1582–1596, 2013.

[7] J. Chandran and A. Joseph. Eye gaze as an indicator for stress level analysis in students. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1588–1593, 2018.

[8] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: A definition, a model, and a knowledge catalogue, 2021.

[9] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[10] L. M. Cysneiros, M. Raffi, and J. C. Sampaio do Prado Leite. Software transparency as a key requirement for self-driving cars. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 382–387, 2018.

[11] R. Desharnais, J. Jobin, C. Côté, L. Lévesque, and G. Godin. Aerobic exercise and the placebo effect: a controlled study. *Psychosomatic medicine*, 55(2):149–154, 1993.

[12] H. Deters, J. Droste, and K. Schneider. A means to what end? evaluating the explainability of software systems using goal-oriented heuristics. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, page 329–338, New York, NY, USA, 2023. Association for Computing Machinery.

[13] H. Deters, J. Droste, and K. Schneider. On the pulse of requirements elicitation: Physiological triggers and explainability needs. In *REFSQ Workshops*, 2024.

[14] J. Droste, H. Deters, M. Obaidi, and K. Schneider. Explanations in everyday software systems: Towards a taxonomy for explainability needs. In *2024 IEEE 32th International Requirements Engineering Conference (RE)*, 2024.

[15] J. Droste, H. Deters, J. Puglisi, and J. Klünder. Designing end-user personas for explainability requirements using mixed methods research. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 129–135, 2023.

[16] S. Fahl. Kapitel 4 Qualitative Methoden. Vorlesung Empirische Methoden (WiSe 2023/24), Leibniz Universität Hannover, 2023.

[17] M. Fechner. Konzept und implementierung einer komponente zur untersuchung des erklärungsbedarfs von software. *Bachelorarbeit, Leibniz Universität Hannover*, 2023.

[18] L. D. Goodwin and W. L. Goodwin. Qualitative vs. quantitative research or qualitative and quantitative research? *Nursing Research*, 33(6):378–384, 1984.

[19] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015.

[20] M. Kuhrmann, P. Tell, J. Klünder, R. Hebig, S. Licorish, and S. MacDonell. Helena stage 2 results. *ResearchGate*, 2018.

[21] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368, 2019.

[22] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[23] M. Lotzkar and J. L. Bottorff. An observational study of the development of a nurse-patient relationship. *Clinical Nursing Research*, 10(3):275–294, 2001. PMID: 11881943.

[24] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[25] L. A. Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, and B. S. Elger. Re-focusing explainability in medicine. *DIGITAL HEALTH*, 2022.

[26] M. Rohs. Kapitel 7 Quantitative Datenauswertung. Vorlesung Empirische Methoden (WiSe 2023/24), Leibniz Universität Hannover, 2023.

[27] T. Speith. How to evaluate explainability? - a case for three criteria. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 92–97, 2022.

[28] C.-H. Tsai and P. Brusilovsky. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 391–396, New York, NY, USA, 2019. Association for Computing Machinery.

[29] University of Science and Technology Beijing USTB. Patent for: "pupil center-corneal reflection (pccr) based sight line evaluation method in sight line tracking system", 2010. available at: `https://patents.google.com/patent/CN102125422A/en`.