Visualization Research Center of the University of Stuttgart (VISUS)

Bachelorarbeit

Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks?

Amer R.

Course of Study:	Informatik
Examiner:	Prof. Dr. Daniel Weiskopf
Supervisor:	Dr. Sandeep Vidyapu, Sita Vriend
Commenced:	July 3, 2022
Completed:	January 3, 2023

Abstract

The designing of an eye-tracking experiment may influence the user behavior on stimuli. Thus, designing a non-intrusive eye-tracking study is essential to understand the user interaction with the stimuli, especially for cognitive-demanding tasks. This thesis explored different alternatives of study design and their impact on the cognitive load of the participant.

Five different designs were posed, four of which presented questions textually with different presentation order of image and question, while one presented questions auditory. A study was conducted where participants performed visual question answering for each of the design and the cognitive load was measured for each design individually. The study revealed that the presentation order of image and question had a significant impact on the cognitive load, while the medium with which the questions were presented, did not.

Kurzfassung

Die Gestaltung eines Eye-Tracking-Experiments kann das Verhalten des Nutzers gegenüber den Stimuli beeinflussen. Daher ist das Design einer nicht-intrusiven Eye-Tracking-Studie wichtig, um die Interaktion des Nutzers mit den Stimuli zu verstehen, insbesondere bei kognitiv anspruchsvollen Aufgaben. In dieser Arbeit wurden verschiedene Alternativen des Studiendesigns und ihre Auswirkungen auf die kognitive Belastung der Teilnehmer untersucht.

Es wurden fünf verschiedene Designs vorgestellt, vier davon präsentierten die Fragen in Textform, wobei die Reihenfolge der Präsentation von Bild und Frage variierte, und ein Design präsentierte die Fragen auditiv. Es wurde eine Studie durchgeführt, bei der die Teilnehmer für jedes Design die Aufgabe des Visual Question Answering (visuelle Frage Beantwortung) durchführten und die kognitive Last für jedes Design einzeln gemessen wurde. Die Studie ergab, dass die Reihenfolge der Präsentation von Bild und Frage einen signifikanten Einfluss auf die kognitive Belastung hatte, während das Medium, mit dem die Fragen präsentiert wurden, keinen hatte.

Contents

List of Figures

List of Tables

Acronyms

AOI area of interests. 13

CL cognitive load. 13

CLT Cognitive load theory. 16

GS6 Guided Search 6.0. 15

NASA-TLX NASA Task Load Index. 25

ROI region of interests. 13

VQA visual question answering. 13

1 Introduction

With sensor technology becoming more available and more convenient to use, eye tracking technology is rising in popularity, not only for research purposes, but also for private usage, as there is a steadily growing market for eye tracking systems. Figure 1.1 shows publications in each year, starting from the year 2000, in which the terms "eye tracking" **OR** "eye-tracking" **OR** "eyetracking" were used in the title and abstract, in `https://app.dimensions.ai`. The figure shows a steady increase of eye tracking in scientific research.

Eye movements are generally made up of saccades and fixations [YS75]. Saccades are fast movements of the eye, which occur between the fixations. Fixations refer to the timeframes in which the eyes focus on a target location. Eye tracking technology has the ability to monitor and record an individual's eye movements and gaze patterns. Using this technology allows for research which can give valuable insights into how humans search and how they interact with different visual inputs. This can be used in a variety of fields, like education [CPS08; VS10] and emotion recognition [LMT20]. Eye tracking can be realized through different approaches [AF13]. Most systems found in the literature are using cameras to perform eye tracking, but there are other approaches like Electrooculography (EOG), which uses electrodes placed on the skin around the eyes to measure the electrical potential generated by eye movements [BWGT10; SPP14]. For eye trackers working with cameras, they typically are divided into two types: Screen-based eye trackers and mobile eye trackers. While screen-based eye trackers are more useful for tracking gaze on a screen, mobile eye trackers can be used to observe gaze on environments. This thesis made use of a screen-based eye tracker, as it was the most suited for the task.

The task participants performed in the study of this thesis was visual question answering (VQA). The term of VQA was proposed in 2015 [AAL+15], and deals with the task of providing an accurate natural language answer to a given image and a natural language question about the image. VQA is used for human studies, as well as for studies using machine learning and artificial intelligence. These studies can give valuable insights into visual search and attention guidance of both humans and machines. Analysis of these studies often use, what is called an area of interests (AOI) or a region of interests (ROI). AOIs are defined regions within a visual stimulus that researchers are interested in studying. In eye tracking research, AOIs are typically defined by the researcher in order to focus the analysis on specific parts of the visual display. Researchers can use AOIs to measure various eye movement metrics, such as the amount of time spent looking at an AOI, the number of fixations made within an AOI, or the order in which different AOIs are visited.

The designing of an eye-tracking experiment may influence the user behavior on stimuli. Thus, designing a non-intrusive eye-tracking study is essential to understand the user interaction with the stimuli, especially for cognitive-demanding tasks. For this thesis, non-intrusive was defined as low in cognitive load (CL). CL refers to the amount of mental effort or information processing that is required by an individual when performing a task or learning new information. That is why CL plays an important role in study design.

1.1 Aim

This study wanted to explored alternatives of study designs and report the recommendations for study design, which are low in cognitive load. The goal is to create a method of collecting data that does not interfere with the natural behavior of the participants. In the case of eye tracking studies, this natural behavior is given by the interaction of participants with the stimuli.

Thesis aim: *Optimize the experimental design in eye tracking studies for cognitive load and report the recommendations.*

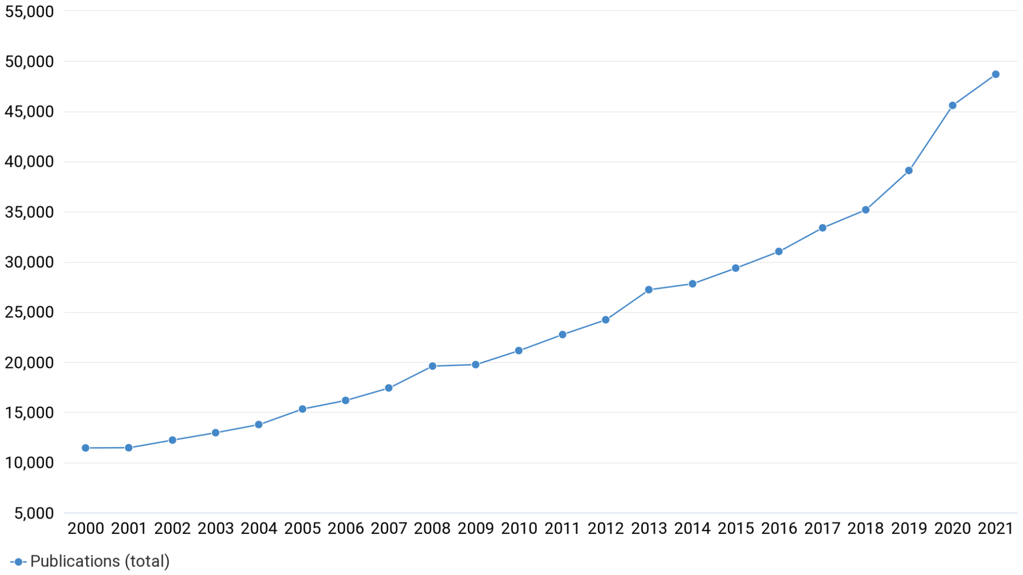


Figure 1.1: Number of publications each year

2 Theory

In this section, the theory of related topics to our research was explored and explained.

2.1 Visual search

When working with eye trackers and eye tracking data, it is important to be familiar with the topic of visual search. Visual search refers to the process of finding a specific item or target within a visual scene. Eye tracking technology has greatly enhanced our understanding of how people perform visual search tasks and has allowed researchers to study the underlying cognitive processes involved. Recent research by JM. Wolfe [Wol21] describes an updated model of visual search: *Guided Search 6.0 (GS6)*. The paper differentiates between two pathways humans use while performing visual search: A non-selective pathway, which includes general information about the stimuli gathered through preattentive processing, and a selective pathway, which allows binding and recognition of objects through selective attention. Furthermore, GS6 specifies five forms of guidance:

1. top-down and
2. bottom-up feature guidance
3. prior attention history
4. value
5. structure and meaning of the scene

Top-down and bottom-up feature guidance use, as the term suggests, features of objects, like color or size, to guide the attention. The literature calls bottom-up feature guidance stimulus-driven, as it refers to the salience a certain object has due to its features. A good example would be how a black dot on a white background would immediately attract attention. Top-down on the other hand is user-driven, meaning the user will guide his attention to specific features, for example, if the user is instructed to look for a red dot, the user will guide his attention to red objects. Prior attention history is about priming and repetition effects, e.g. if the last task required a blue object, it is more likely that the attention will be guided to blue the next time. Guidance through value means valuable features will be prioritized. The paper explains the value of features in terms of rewarded features and punished features. Structure and meaning of the scene refers to guidance through recognizing properties of the scene, e.g. you will not search for a cat on the ceiling. This example shows that it does not have anything to do with the properties of the cat, but more so with the ceiling, which is part of the scene.

2.2 Cognitive load theory

Cognitive load theory (CLT) is a theory that focuses on the role of CL in the learning process. It was developed by J. Sweller and his colleagues in 1988 [Swe11; Swe88] and has since become a widely recognized and influential theory. According to CLT, the human brain has a limited capacity for information processing and working memory. When we are presented with new information or try to learn something, we must process this information in our working memory, which is a temporary storage system that allows us to work with the information. If the CL of the task or material is too high, it can exceed the capacity of our working memory, leading to decreased learning and performance.

CLT suggests that there are three types of CL:

- Intrinsic CL: This is the inherent complexity of the material being learned or the task being performed. It is related to the structure and organization of the information and cannot be changed.
- Extraneous CL: This is the redundant information or task demands that do not contribute to learning or performance. It can be reduced or eliminated through good instructional design.
- Germane CL: This is the mental effort required to integrate new information with existing knowledge and make connections between them. It is necessary for learning and can be increased through the use of appropriate instructional strategies.

Even more important for this thesis, was establishing methods for measuring CL. B. Morris et al. [MDG14] differentiate between 3 categories:

- Indirect measures: These involve performance indicators of participants measured during task or learning activity, and using these measurements as an indicator of CL. Examples of such measurements would be speed or error rate. Indirect measures can provide objective data on CL, but they may be influenced by other factors, such as skill level or fatigue.
- Subjective measures: These are gathered by asking the participants about their own CL during a task or learning activity, which can be done through verbal or written responses, such as questionnaires. It is important to note, however, that these might not always be entirely accurate, as subjective measures might get influenced by other factors, like the mood of the participant.
- Direct measures: There are two approaches to direct measurement of cognitive load. The first one involves physiological measurements, like heart rate or pupillary response. The second approach for direct measuring of CL is using dual task, meaning that an additional task must be performed in addition to the primary task. Depending on the CL of the primary task, the performance in the additional tasks will suffer accordingly.

It is important to consider the strengths and limitations of each method when choosing a method to measure cognitive load. In some cases, it may be useful to use multiple methods to get a more complete picture of cognitive load.

2.3 Eye tracking studies

A literature research was done to identify and collate the relevant works, study designs, and questionnaire. As the goal of this thesis is to explore different alternatives of study designs and analyze these designs in terms of CL, previous eye tracking studies were explored and their structure and procedure were analyzed.

S. Chen et al. [CJYZ20] recently published a paper about VQA. The aim of their work was to analyze attention of both humans and machines, through the use of experiments and in case of the human analysis, with help of eye tracking. While the objective of this thesis was a different one, the study design of this thesis was still closely related to theirs. The images and questions were taken of the same underlying dataset[HM19], but the selection process was a different one, as the study of this thesis needed only 100 question-image pairs, while they selected 987 images and 1422 questions. For presentation of questions and images, their study served as a basis for our study: Participants were able to see the question before and after the image for an unlimited amount of time, images on the other hand were only visible for three seconds.

Other research, which is related to ours, was done by G. Underwood et al. [UJR04]. They conducted two separate eye tracking experiments, with the goal of analyzing how participants compare information of a text and a graphic component. The task was a verification task, where they had to label sentences with either true or false. For the first experiment, they showed text and graphic component at the same time. In the second experiment, however, they separated the components. Furthermore, half of the participants saw the text component first, while the other half saw the graphic component first. This design difference will be tested in the study of this thesis as well.

3 Method

This section aimed to provide a clear and detailed description of the research methods that were used in the study.

3.1 Research questions

With the aim of this thesis in mind, the study wanted to answer the following research questions:

1. Does the presentation order of image and question significantly impact CL in an eye tracking study?
2. Does presenting questions auditory, rather than displaying them as text, affect CL in an eye tracking study?

3.2 Hypotheses

With these research questions in mind, the following hypotheses were posed:

Hypothesis H_1: Participants rated Design 1 a lower CL compared to Design 4

Hypothesis H_2: Participants rated Design 1 a lower CL compared to Design 5

Hypothesis H_3: Participants had a higher accuracy during Design 1 compared to Design 4

Hypothesis H_4: Participants had a higher accuracy during Design 1 compared to Design 5

3.3 Participants

Demographics were recorded through a pre-test survey (appendix B), which was completed by all the participants.

15 participants signed up for the study, and of these 15, 13 actually participated in the study.

Out of all the participants, ten were male, two were female, and one participant specified other gender.

Participants with full colorblindness were excluded from this study, as it contains questions, which require the participant to differentiate between colors. Participants with Red-green colorblindness or Blue-yellow colorblindness were allowed to participate, because these types of colorblindness do not affect the study significantly, as this was tested beforehand. Out of all participants, two were red-green colorblind, the rest of the participants were not colorblind. Six out of 13 participants had corrected eyesight.

Participants had to be at least 18 years old to participate in this study. Participants were between the age of 18 and 29.

Seven participants studied English for 11 or more years in school, five studied it for 6 to 10 years, and only one participant for 5 or fewer years. Ten participants were very confident in their English, while the remaining three were a little confident.

All the participants dominant reading directionality was "Left to right"

Every participant except for one finished high school as their highest degree or level of education. One participant finished a bachelor's degree.

3.4 Apparatus

The Eye tracker used in this study is the **"Tobii Pro Spectrum"**, which is a screen-based eye-tracker manufactured by the company "Tobii". It consists of an eye tracker unit and a monitor. The eye tracker unit can be used with the provided monitor, but it is also possible to operate without it. The monitor had a diagonal of 24", a resolution of 1920x1080 and a refresh rate of 60Hz. This was sufficient for the study, which is why it was used and not swapped out. Eye tracking was performed for both eyes and with a sample frequency of 600Hz.

Participants used a head rest during the whole study, which was mounted to the desk at a distance of approximately 63 cm away from the screen.

Giving answers was entirely implemented through dialogue boxes with drop down menus, which allowed the participant to only use the mouse as an input device.

3.5 Stimuli

The stimuli and questions used in the study are from the GQA Dataset[HM19]. The GQA Dataset is a VQA Dataset build to fix shortcomings of its predecessors. It consists of real-world images and uses scene graphs of these images to generate questions. Figure 3.2 shows an example image of the dataset. Text to speech was used to generate audio files for the questions in Design 5.

3.5.1 Manual quality control of images and questions

As the GQA Dataset consists of 113 018 images and 22 669 678 questions, which were generated through a question engine, not all of these were fitting for this study. To ensure quality and suitability of images and questions for this study, manual quality control was performed. The criteria used for this quality control were:

- All images used in the study had to have the same resolution. For this study, images with a resolution of 1024x768 were chosen, as it was the highest resolution with a great number of images available.
- Ambiguous questions were discarded, e.g. the question *"On which side stands the man"* to an image in which there is a man standing on the left and another one on the right
- In the dataset, answers were also provided for each question. In some cases, these answers were false. All questions with false answers were discarded.
- Questions which we were not able to answer were discarded. E.g. *"On which side stands the man"* was the question, and we were not able to find a man in the picture.
- Questions where you can make an educated guess were discarded. E.g. *"What color is the sky"* is the question, and *"Blue"* would be the answer.

As already established in section 2.2, CL during the task and the difficulty of the task are strongly interdependent. Therefore, it is necessary to keep the difficulty of the image-question-pairs of each design as consistent as possible. This is obviously a difficult task by itself, as measurements of difficulty can be limited and highly subjective.

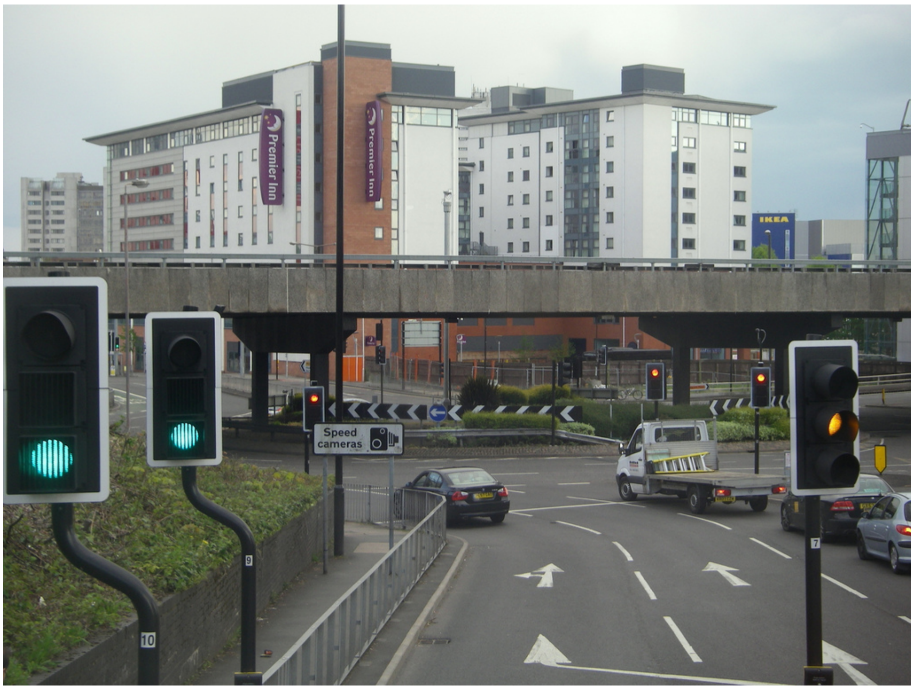


Figure 3.1: Example of an image used for the study

3.5.2 Presentation of stimuli

As the monitor was capable of displaying a resolution of 1920x1080, images were upscaled to a resolution of 1440x1080 and centered in the middle of the screen, while the remaining screen was filled with a gray (Color-hex:#7f7f7f) background. Figure 3.2 shows an example of how images were presented in the study. This particular image is from the example tasks at the beginning of the study.

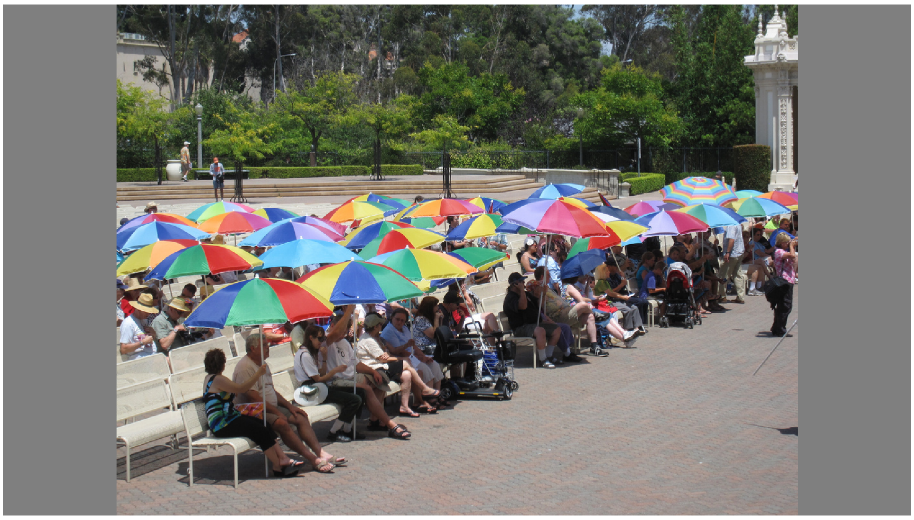
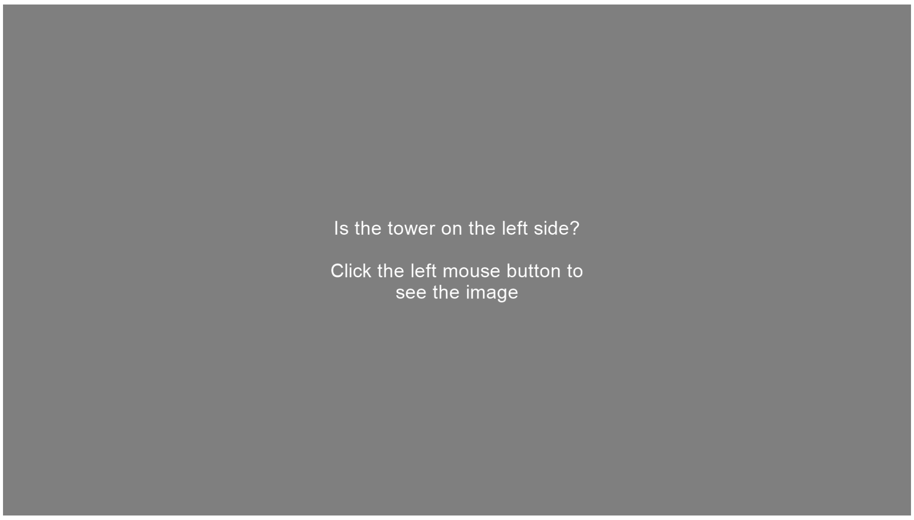


Figure 3.2: Example of how the images were presented

Questions were either displayed as white (Color-hex:#ffffff) text on a gray (Color-hex:#7f7f7f) background, or were played as sound through headphones. Figure 3.3 shows an example of how questions and instructions were presented in the study. This particular question is from the example tasks at the beginning of the study. Examples for each design can be found in the appendix C.



Is the tower on the left side?

Click the left mouse button to see the image

Figure 3.3: Example of how the questions were presented

4 Data Collection

This section defined what kind of data was collected and explained the reasoning behind it.

4.1 Pre-test survey

Each participant filled out a pre-test survey (appendix B) before the actual study started. The survey was completed digitally on the computer, using a dialogue box in the same program the study ran in. This pre-test survey had the purpose of gathering demographic data and give an overview of the participants. The questions asked in the pre-test survey were about age, gender, how long they studied the English language and how confident they were in it, reading directionality, the highest degree they completed, if they have corrected eyesight and if they are colorblind.

4.2 Independent Variables

The independent variables used for this study were:

- Question medium.
- Order of stimuli and question.

For the study, every participant tested all independent variables, which were implemented in 5 different designs. The idea of testing between subjects was also considered, but as this would not have acquired enough data with the number of participants, the idea was discarded.

4.2.1 Question medium

As a reminder, the task posed to the participants was VQA, meaning, given an image, they had to give an accurate natural language answer to a natural language question about the image.

The dependent variable refered to the medium with which the tasks or rather the questions were delivered to the participants. There were two different media used for delivering the questions to the participants:

As text displayed on the screen or *as audio played by headphones*.

If the medium was audio, the question was played once at the start and the participant was able to replay the audio once by pressing the left mouse button. These instructions were also shown as text displayed on screen, during each trial. After that, the participant continued by pressing the left mouse button again (this instruction was also displayed on the screen) and the image was shown for the fixed time of three seconds. After the image was shown, the participant was able to hear the question one more time. The participant was able to continue and answer the question by pressing the left mouse button.

4.2.2 Order of stimuli and question

If the question was displayed on the screen, there were 4 different orders in which the image and question were shown, while the image was always shown for a fixed duration of 3 seconds and the question for as long as the participants did not press the mouse to continue:

- Question, image, question
- Question, image
- Image, question, image
- Image, question

With that, there were 5 different designs in total:

Design List		
Design ID	Question medium	Order
1	Text	Question, image, question
2	Text	Question, image
3	Text	Image, question, image
4	Text	Image, question
5	Audio	Possible to listen to at any time

Table 4.1: Table of different design setups.

In this thesis, designs were referred to by their ID.

4.2.3 Order of the designs/conditions

To counteract carryover effects [Bro12], which are changes in a participant's response that can occur simply because of the order in which the experimental conditions are presented, counterbalancing was performed. The term counterbalancing refers to the usage of different orders of presentation for different participants in an experiment.

To perform full counterbalancing, one would need to go through each possible order of conditions. This is the most effective way to control carryover effects [DB14], because it ensures that each condition is presented in every possible order. Full counterbalance of five different conditions would generate 5! = 120 different orders of conditions, which would need at least 5! = 120 participants.

As there were not nearly enough participants in this study, partial counterbalancing was used instead. Partial counterbalancing refers to a situation where only some of the possible orders of presentation are used. This can help to control carryover effects, but it is not as effective as full counterbalancing. For partial counterbalancing, one only has to ensure that each condition is represented at each position in the order once. With that, five conditions will give us five different orders.

One possibility, to implement partial counterbalancing, is using Latin squares. A Latin square is a square grid in which each cell contains one of a set of symbols, and no symbol is repeated in any row or column.

1	2	3	4	5
2	3	5	1	4
3	4	2	5	1
5	1	4	2	3
4	5	1	3	2

Table 4.2: Latin square used for counterbalancing.

Each number representing the design ID and each column representing the order, staring from the top. With that, the five different orders of presentation used for the study were:

1. Design 1 → Design 2 → Design 3 → Design 5 → Design 4
2. Design 2 → Design 3 → Design 4 → Design 1 → Design 5
3. Design 3 → Design 5 → Design 2 → Design 4 → Design 1
4. Design 4 → Design 1 → Design 5 → Design 2 → Design 3
5. Design 5 → Design 4 → Design 1 → Design 3 → Design 2

4.3 Dependent Variables

The dependent variables used for this study were:

- NASA Task Load Index (NASA-TLX)
- Accuracy
- Gaze data

NASA-TLX and accuracy were used for hypotheses testing, while the gaze data was used for an exploratory data analysis.

The dependent variables were measured the same way for every participant. Data recording started at the calibration of the eye tracker. That means, the three example questions were not recorded, and their only purpose was for the participants to get an understanding of the tasks they had to perform. Instructions also made it clear to the participant that the example questions were not being recorded.

4.3.1 NASA-TLX

To measure the cognitive load in this study, the NASA-TLX was used as the main measurement in this study (appendix D), as it is highly established in scientific research[Har06].

“NASA-TLX is a subjective workload assessment tool. NASA-TLX allows users to perform subjective workload assessments on operator(s) working with various human-machine systems.” “the NASA Task Load Index (NASA-TLX) consists of six sub-scales that represent somewhat independent clusters of variables: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance.”[20]

For this study, the sub-scale for Physical Demands was cut, as the experiment did not require the participants to perform physical demanding tasks. For the other five sub-scales, 7-point scales were used, as they were sufficient for this study.

The description of each sub-scale was taken from the appendix section in [Har06] and shortened before being used in the study. The five sub-scales with their respective description were:

1. Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?
2. Temporal Demand: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?
3. Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
4. Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
5. Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter?

For *Mental demand, Temporal Demand, Effort* and *Frustration* the options to choose from were:

Very Low, Low, Somewhat Low, Neutral, Somewhat High, High, Very High

For *Performance*, it was:

Very Good, Good, Somewhat Good, Neutral, Somewhat Poor, Poor, Very Poor

To be able to work with the NASA-TLX data, each option from the questionnaire was assigned to a value (see figure 4.3). With these values established, the cognitive load could now be defined. To calculate the cognitive load, the value of each subscale got multiplied with its weight. There are multiple ways to set these weights [VMKH21]. This thesis decided to use the same weight for each subscale value, as this approach allowed for as little interference into the study as possible and was shown to be valid by previous research [DN08; MBE95; Nyg91]. As five subscales were used, the weight for each one was $\frac{1}{5}$, and with that, the cognitive load was given by:

$$\text{Cognitive load} = \frac{\text{Mental Demand} + \text{Temporal Demand} + \text{Effort} + \text{Frustration} + \text{Performance}}{5} \tag{4.1}$$

Options	Values
“Very Low” OR “Very Good”	1
“Low” OR “Good”	2
“Somewhat Low” OR “Somewhat Good”	3
“Neutral”	4
“Somewhat High” OR “Somewhat Poor”	5
“High” OR “Poor”	6
“Very High” OR “Very Poor”	7

Table 4.3: Values assigned to each option from the questionnaire.

4.3.2 Accuracy

For each question the participant answered, the answer got recorded. The method for collecting answers was a dialogue box with a drop-down menu (see figure 4.1) and was the same for all designs. To counter typing mistakes, the participants were able to choose between two possible answers and the third option, which was *“I am not sure”*. As most of the questions were binary (Yes-no-questions, Left-right-questions, Top-bottom-questions, etc.) it does not impact the answers for these questions. For all other questions, it arguably made them easier, which in turn reduced cognitive load for all non-binary-questions. Also, it was possible for the participant to simply guess one of the two options instead of picking *“I am not sure”*, which has to be taken into consideration during the data analysis.

With the answers given by the participants, the accuracy, which refers to giving right answers to the given questions and images, was calculated for each design individually. The accuracy got calculated by taking the number of all the right answers and dividing it by the number of questions:

$$(4.2)\quad \text{Accuracy} = \frac{\text{Number of right answers}}{\text{Number of questions}}$$

A high accuracy implies a low error rate, and a low accuracy implies a high error rate. The error rate is an indirect measure of cognitive load, as shown in previous research [AS90; Ayr01; MDG14], meaning accuracy is an indirect measure of cognitive load as well.

4.3.3 Gaze data

After the calibration of the eye tracker, gaze data of the participant was recorded for the remainder of the study. This gaze data contained:

- Timestamp: Provides time stamps for system and device.
- Gaze origin: Describes position where the gaze vector starts.
- Gaze point: Describes the position of the intersection between gaze and screen.
- Pupil diameter: Calculated size of the pupil by registering images of the eyes.

Gaze data were recorded for both eyes and with a sample rate of 600Hz.

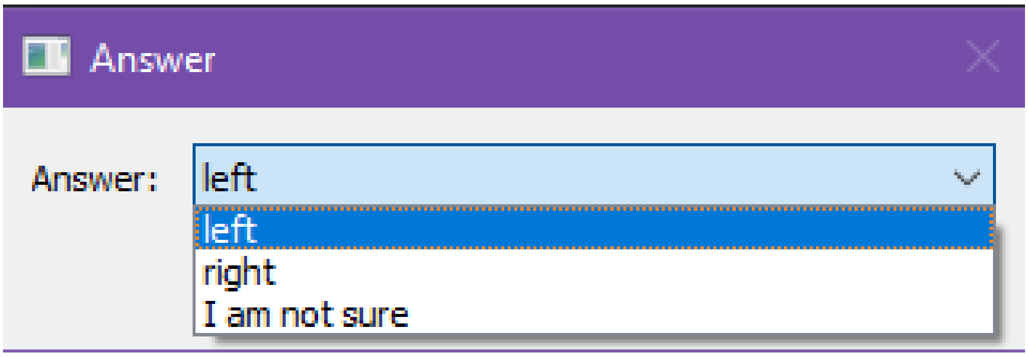
Pupil diameter

Measurements of the pupil diameter can also be a metric for cognitive load [KDN+18], or in a more general sense, a direct measure of mental activity [HP64]. However, pupil diameter is not a reliable measurement, as it is heavily influenced by outside factors, like emotion, stress, and pain, as well as the stimuli used. To get a somewhat reliable measure, one would have to use stimuli which are consistent in size, color, and luminance [SBS+22]. This is why this measurement was not used for this study.

Fixations

Fixations were extracted from the gaze data using the I2MC algorithm [HNKH17]. This algorithm used gaze data as an input and outputted a list with all the fixations of the data. The list included information of the fixation about start and end, duration and coordinates and assigned each fixation to the respective trial and participant.

Fixations were used as part of the exploratory data analysis.



Answer

Answer: left

Figure 4.1: Example of a dialogue box for answer input

5 Experimental study

5.1 Pilot studies

Two pilot studies were conducted before the final study. These two pilot studies had different goals, and that is why they used different versions of the experiment.

5.1.1 First pilot study

The first pilot study was conducted with two participants, which knew about the study and its rough procedure beforehand. It used a shortened version of the experiment where each design block consisted of only three question-image pairs, instead of the 20 in the final experiment.

The main goals of this first pilot study were to get opinions on the overall procedure and implementation of the study, as well as identifying unclear instructions.

The pre-test survey was found to be missing some more demographic questions like e.g. if the participant was colorblind (and if so, then what kind of colorblindness) and also the dominant reading directionality. These were added to the existing questions. Some of the existing question got some additional options added as well.

In some parts of the study, instructions were missing or unclear, and also some minor spelling mistakes were found.

The pilot study used a version, where it was possible to replay the sound in Design 5 at any point. For a non-apparent reason, this caused the program to crash and thus prevented the participant from continuing the study. After the pilot study, a few different approaches were tried to fix this issue, but ultimately this was not fixed, which is why different approaches were tried, and fixed audio play was implemented in the final version of the program.

5.1.2 Second pilot study

The second pilot study was conducted with one participant, who knew nothing about the experiment beforehand, except for it being an eye-tracking study. It used the whole blocks with 20 question-image-pairs for each design. This was necessary, as this pilot study should be as close to the real experiment as possible.

The main goals now were to get a good estimation of the time needed to perform the study, to see if everything works properly, and to make a rough script for myself.

The time measured for this pilot study included the time for all the preparations, these included reading through the study description and consent form, signing the consent form, sanitizing of equipment and adjustment of the head mount, As well as the time for the actual study. The measured time was about 35 minutes. This measurement was used as the estimated time for the study in the study description, as well as the invitations for the study.

Like in the first pilot study, the participant pointed out unclear or missing instructions, which were fixed after.

5.2 Software

5.2.1 Software for study implementation

When working with a Tobii eye-tracker, it seems reasonable to use official software from Tobii for the implementation of the Experiment. While trying Tobii's official software, Tobii Pro Lab, it became apparent that implementation with this software would not be sufficient for this thesis.

It is a great tool for recording and visualizing eye-tracking data, but it does not provide enough functionality for the different experimental designs of this study. For example, working with sound, which is essential for this thesis, seems to be very laborious. Another possibility for capturing eye-tracking data with a Tobii Pro Spectrum is Psychopy. Psychopy offers enough study design tools, which allow implementation of the experimental designs of this thesis. Ultimately, **"Titta"** was used for implementation: "an open-source toolbox for controlling eye trackers manufactured by Tobii AB from MATLAB and Python" [NAN20], as it provides a working interface for the "Tobii Pro Spectrum" in Python. "Titta" was used to implement the pre-test survey (appendix B) and the eye tracking experiment, which included calibration of the eye tracker, recording of gaze data, presentation of images and questions, recording of answers and recording answers of the questionnaires. The source code can be found at `https://github.com/Amcee/Titta-Eyetracking-Experiment`. The file **study_final.py** runs the study.

5.2.2 Software for data analysis

For the data analysis, mostly custom written Jupyter notebooks were used, as they provide options for easy calculation of statistics and statistical tests, as well as quick visualization of data. They were used for formatting data, calculating statistics and statistical tests and plotting boxplots. They can be found at `https://github.com/Amcee/DataAnalysisBT`.

The exploratory gaze tracking data analysis was done with a new toolkit, called WebVeta[1], a web-based visual eye tracking analytics toolkit that could be integrated into gaze data analysis workflow. The toolkit supports labelling fixations with AOIs, a multifaceted combination of AOI and non-AOI comparative metrics and data filtering options to visually explore gaze patterns across samples, space and time. The visualizations of gaze data in this thesis were created using WebVeta.

[1] The URL to the toolkit is anonymized as the work is unpublished and currently under peer review

5.3 Procedure

5.3.1 Preparations

The study was conducted at the "Projektlabor 2" at the Visualization Research Center (VISUS) of the University of Stuttgart. To keep the illumination inside the room as consistent as possible for all participants, light from the outside was blocked using the curtains, and the lights inside the room were used instead. After the participant entered the lab and were seated, they were immediately informed that they are able to ask questions at any times. Then, the study description and consent form were handed to the participant, and they got time to read through both. While the participant was reading, the researcher sanitized the head mount, headphones and the computer mouse. To make sure everything was clear, the participant and the researcher went through the procedure part of the study description together. The Researcher then explained to the participant how to adjust the head mount and asked the participant to adjust it for themselves. In the end, the researcher mentioned that there will be instructions throughout the whole study and asked if there are any questions. If there were no questions or after the remaining questions were answered, the researcher started the program, which might take a few seconds.

5.3.2 Main study

Once preparations were finished, the main study starts with the pre-test survey (appendix B), in which participants answer demographic questions about themselves.

When the participant completed the pre-test survey, general information about the study will be displayed. The next step was a training section. The training section consist of three image-question pairs, where the question will be displayed before and after the image.

After the training section, the calibration of the eye tracker and recording starts. Once calibration was finished, the participant completed each design in the specified order, depending on his participant ID.

After finishing all designs, the program exits and the researcher thanks the participant for their participation.

5.4 Results

First, descriptive statistics of both NASA-TLX and accuracy were calculated and visualized, with the purpose of providing information about the data sets. After that, the hypotheses were tested. Lastly, an exploratory data analysis was performed using the recorded gaze data.

5.4.1 Excluded data

During one study, there were hardware difficulties regarding the headphones, which made it impossible for the participant to answer questions from Design 5. That is why for this participant no data was recorded for Design 5 and hence cannot be used. The Wilcoxon signed-rank test requires dependent samples, which is why the data from that same participant for Design 1 was not used when testing H_2 and H_4.

5.4.2 Preparations

$Cronbach's\ \alpha$ was calculated to ensure the internal consistency of the NASA-TLX data. Using all questionnaire answers available, a value of 0.91 for $Cronbach's\ \alpha$ was calculated. Values > 0.9 can be interpreted as excellent [GG03]. This means the internal consistency of the different sub-scales in our modified NASA-TLX was assured, and the data can be used for further analysis.

With a valid internal consistency, then the CL was calculated, using equation 4.1. This was done using a Jupyter notebook, where the questionnaire answers were first converted to values, and then used the values to calculate the CL.

5.4.3 Descriptive statistics CL

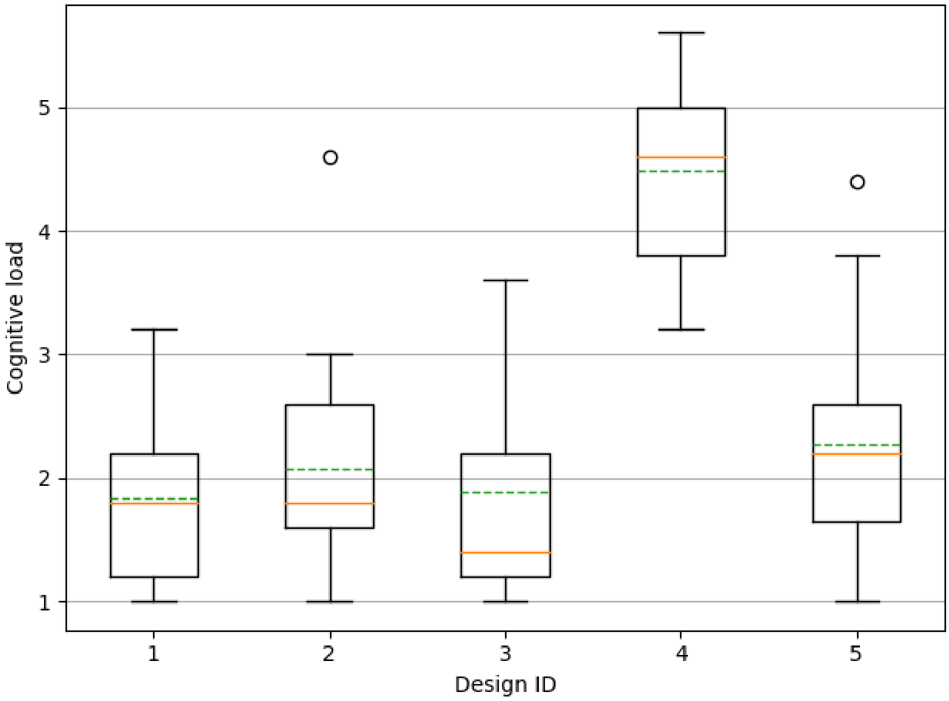


Figure 5.1: Boxplots of cognitive load for each design

Visualized of the calculated values was realized by boxplots as seen in figure 5.1. The main body of the boxplot shows the quartiles and the median's confidence intervals. Horizontal lines inside each box show the median. The vertical lines extending to the most extreme, non-outlier data points. Dashed lines represent the means and the circles represent outliers.

Table 5.1 shows values for the mean, range and variance of the CL, allowing for quick comparisons.

The possible values for the CL, were defined through table 4.3 and the equation 4.1. The lowest value possible was 1, the highest was 7. As already mentioned in section 2.2, a too high CL can lead to decreased performance. That is why a low CL was desired.

Design ID	Mean	Range	Variance
1	1.83	2.20	0.57
2	2.08	3.60	0.90
3	1.89	2.60	0.87
4	4.49	2.40	0.58
5	2.27	3.40	1.12

Table 5.1: Overview of descriptive statistics for the cognitive load

The data suggested that design 4 differs the most from all other designs in terms of CL. Design 1 and 3 showed a lot of similarities in mean, quartiles and extremes. If a ranking according to this data was to be suggested, and with the goal of a low CL in mind, Design 1 and 3 would have the lowest CL, Design 2 would have the second lowest CL, followed by Design 5 in third, and Design 4 in last place.

5.4.4 Descriptive statistics accuracy

Further examination was performed using the accuracy participants achieved for each design individually. The first step was calculating the accuracy using the answers of the participant and 4.2. After that, visualization of the accuracy data was done by using boxplots as seen in figure 5.2, and the values of mean, range and variance can be found in table 5.2.

Possible values for accuracy range from 1 to 0, 1 meaning all answers were correct, 0 meaning no answer was correct. A high accuracy was desired, as it implies a low error rate.

Design ID	Mean	Range	Variance
1	0.95	0.15	0.27e-02
2	0.93	0.15	0.18e-02
3	0.92	0.20	0.42e-02
4	0.64	0.55	2.09e-02
5	0.85	0.35	0.11e-02

Table 5.2: Overview of descriptive statistics for the accuracy

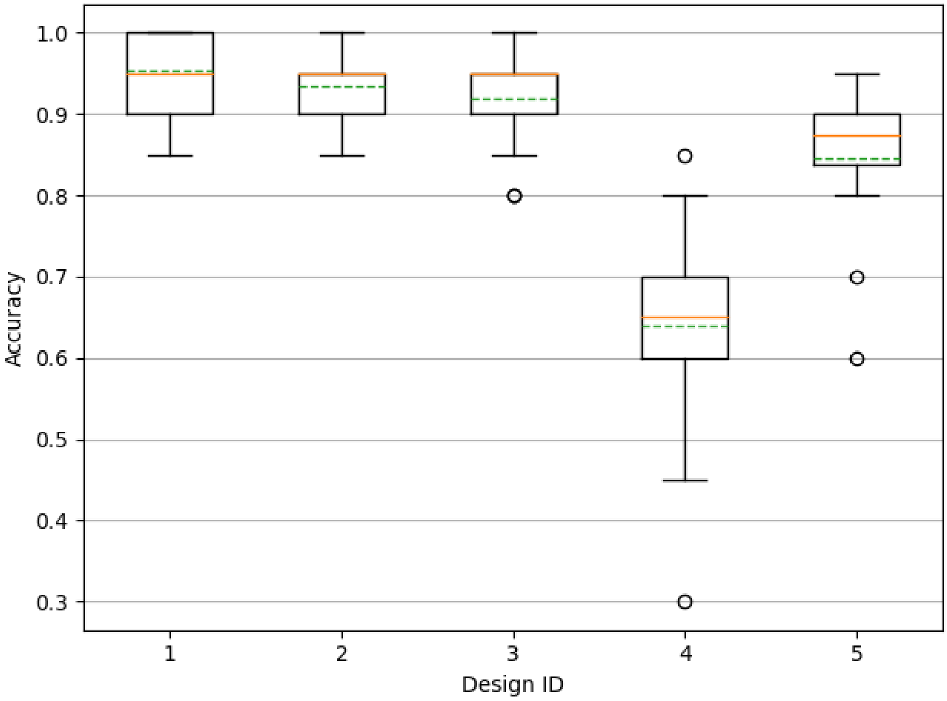


Figure 5.2: Boxplots of accuracy for each design

As before, design 4 stood out from the rest. Compared to CL, Designs 1, 2 and 3 are more similar in accuracy. Again, a ranking according to this data was posed. Designs 1, 2 and 3 were similar enough in accuracy to share the first place, followed somewhat closely by Design 5 in second place, and like before, Design 4 in last place.

5.4.5 Data analysis CL

To check if the CL is normally distributed, a Shapiro-Wilk test [SW65] was performed, which showed that the distribution of the CL departed significantly from normality (W = 0.90, p < .001).

To test the hypotheses, H_1 :

Participants rated Design 1 a lower CL compared to Design 4,

and H_2 :

Participants rated Design 1 a lower CL compared to Design 5,

two separate Wilcoxon signed-rank tests [Woo07] were performed. The Wilcoxon signed-rank test was chosen, because the hypotheses compared to dependent groups, and the distribution of the data did depart significantly from normality.

The first Wilcoxon signed-rank test indicated that the CL of design 4 was statistically significantly higher than the CL of design 1 (W = 0.0, p < .001).

The second Wilcoxon signed-rank test indicated that there was no statistical significant difference in CL between Design 1 and Design 5 (W = 19.0, p = .21).

Dependent variable	NASA-TLX		
Statistical test	Shapiro-Wilk	Wilcoxon signed-rank	Wilcoxon signed-rank
Groups	ALL	1 and 4	1 and 5
test statistic	W = 0.90	W = 0.0	W = 19.0
p value	p = < .001	p = < .001	p = .21

Table 5.3: Overview of test statistics for cognitive load

5.4.6 Data analysis accuracy

For this section, the same steps of the previous section were followed, but with the calculated accuracy of each design instead of the cognitive load.

So the first step was to perform a Shapiro-Wilk test. It showed that the distribution of the accuracy departed significantly from normality (W = 0.82, p < .001).

Again, two separate Wilcoxon signed-rank test were performed, this time for H_3 :

Participants had a higher accuracy during Design 1 compared to Design 4,

and for H_4 :

Participants had a higher accuracy during Design 1 compared to Design 5.

The first Wilcoxon signed-rank test indicated that the accuracy of Design 1 was statistically significantly higher than the CL of design 1 (W = 0.0, p < .001).

The second Wilcoxon signed-rank test indicated that the accuracy of Design 1 was statistically significantly higher than the CL of Design 5 (W = 19.0, p = .21).

Dependent variable	Accuracy		
Statistical test	Shapiro-Wilk	Wilcoxon signed-rank	Wilcoxon signed-rank
Designs	ALL	1 and 4	1 and 5
test statistic	W = 0.82	W = 0.0	W = 0.0
p value	p < .001	p < .001	p = .01

Table 5.4: Overview of test statistics for accuracy

5.5 Discussion

The hypothesis were:

Hypothesis H_1: Participants rated Design 1 a lower CL compared to Design 4

Hypothesis H_2: Participants rated Design 1 a lower CL compared to Design 5

Hypothesis H_3: Participants had a higher accuracy during Design 1 compared to Design 4

Hypothesis H_4: Participants had a higher accuracy during Design 1 compared to Design 5

H_1 and H_3 investigated the influence of the presentation order of image and question on the CL of the eye tracking study. Both hypotheses were supported by the results. The results suggested that there was a significant difference in CL and in accuracy depending on the presentation order. This was also supported by the descriptive statistics of both variables. Designs where the image was shown after the participants had knowledge of the question, showed lower cognitive load for these designs. This can be explained by top-down feature guidance. For the participants to be guided by the features of the searched for object, they would have to get information about the object before seeing the image, to be able to identify specific features and use these as guidance.

H_2 and H_4 investigated the influence of the medium of the question on the CL of the eye tracking study. While H_4 was supported by the results, H_2 was not. The results suggested that there was a significant difference in accuracy depending on whether the question was presented textually or auditory, but no significant difference in CL. This different outcomes for CL and accuracy might be due to influence of guessing, which could falsify the accuracy measure to some extent. Still, the descriptive statistics for the CL showed a slightly higher CL for Design 5, compared to Design 1, 2 and 3, indicated through a higher mean, median and variance.

Through these results, it was possible to provide answers to the research questions:

Does the presentation order of image and question significantly impact CL in an eye tracking study?

Yes, it does.

Does presenting questions auditory, rather than displaying them as text, affect CL in an eye tracking study?

No, it does not.

Considering all of these results, the recommendation this study suggests is presenting questions before images and presenting them textually instead of auditory.

5.6 Exploratory data analysis

This section analyzed gaze data collected during the study and aimed to find connections between gaze data and the results of the previous section. For the exploratory data analysis, fixations of the gaze data were extracted first, to compare number and duration across designs. After that, the gaze data was analyzed using WebVeta (see section 5.2.2). Gaze data was analyzed for images exclusively.

5.6.1 Fixations

Table 5.5 shows the descriptive statistics of the number and duration of fixations. The values for number of fixations were measured for each image, while the values for the duration were measured for each fixation.

	Number of fixations per image			Duration of fixations in ms		
Design ID	Mean	Range	Variance	Mean	Range	Variance
1	10.73	10	3.67	244.19	1568.33	17124.39
2	10.77	11	3.49	244.20	2370.01	22906.07
3	10.56	10.5	2.77	247.55	2576.67	25811.91
4	11.32	10	3.27	232.30	1310.00	12769.17
5	10.20	11	4.23	260.24	1800.00	27237.08

Table 5.5: Overview of descriptive statistics for the fixations

A established in the previous section 5.5, Design 4 had the highest CL. Design 4 had the highest average number of fixations. As the images were always shown for a fixed 3 seconds, the average duration is also the smallest for Design 4. This result matches the result of previous research [UJR04], which was mentioned in section 2.3.

5.7 Analysis using WebVeta

For the analysis using WebVeta, one sample of Design 1, 4 and 5 was analyzed, as an analysis of all images would not be possible in terms of time. For each sample, there was a AOI created for the desired object of the question, and these AOIs were used to calculate how percent of the fixations were on the desired object. While this method was not optimal and had its flaws, it still gave some insights.

Before the data was useable in WebVeta it first had to be converted into the right format. This was done with a Jupyter Notebook (see section 5.2.2).

Figure 5.3 shows the image and AOI for the sample of Design 1. The question, which participants had to answer, was: *"On which side is the player?"* For Design 1 the analysis showed that 22% of all fixations hit the AOI. Figure 5.4 shows the percentage of fixations that hit the AOI for every participant, with colors closer to white indicating a higher value, and colors closer to back indicating a lower value. The matrix shows that every participant hit the AOI, even though percentages are different.

Design 4 had a way lower percentage, with only 3% of fixations hitting the AOI. Figure 5.5 shows image and AOI for the sample, and the question was: *"Is the steering wheel black or white?"*. The matrix in figure 5.6 shows that only four out of all 13 participants the AOI was hit at all.

Lastly, figure 5.7 shows image and AOI of the sample for Design 5. The question was given by: *"Which side is the black bag on, the right or the left?"* The percentage of all fixations that hit the AOI was 15% for this sample, and eight out of 12 participants were able to score a percentage above 0.

There is a clear distinction for Design 4 and the other two, in overall percentage of all fixations that hit the AOI, as well as the number of participants that hit the AOI at least once. This result supports the previous findings discussed in section 5.5.

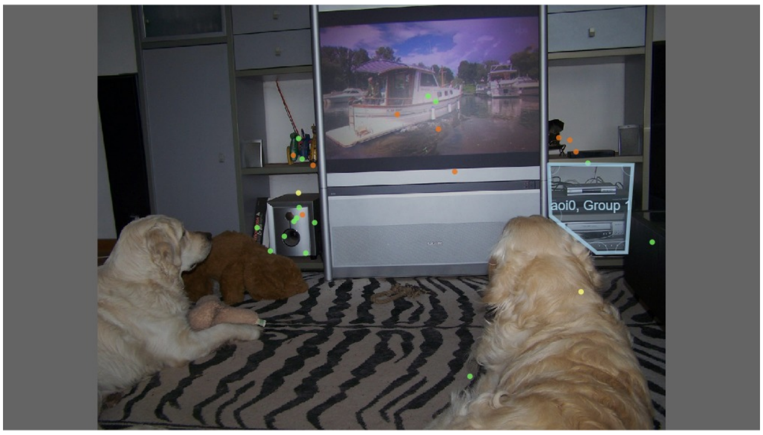
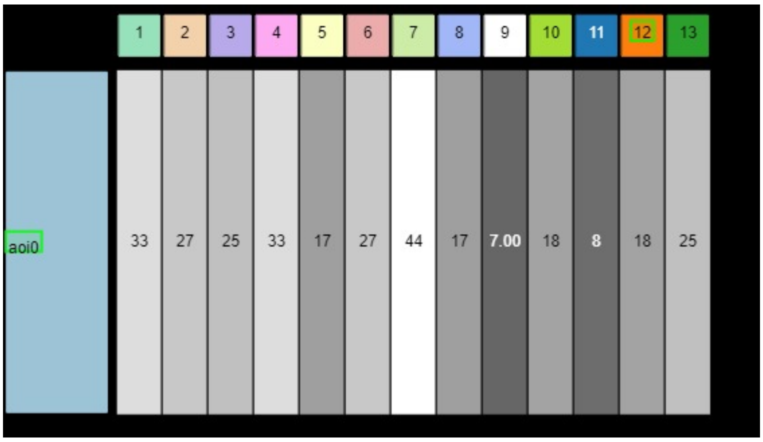


Figure 5.3: Image and AOI for the sample of Design 1



	1	2	3	4	5	6	7	8	9	10	11	12	13
aoi0	33	27	25	33	17	27	44	17	7.00	18	8	18	25

Figure 5.4: Matrix showing the percentage of fixations that hit the AOI for the sample of Design 1

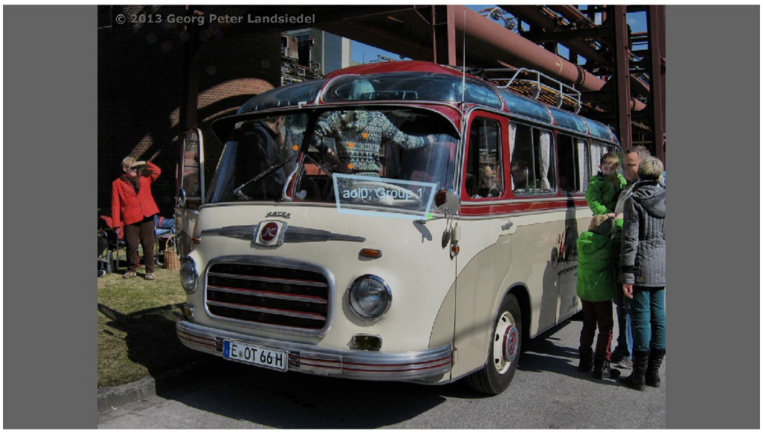
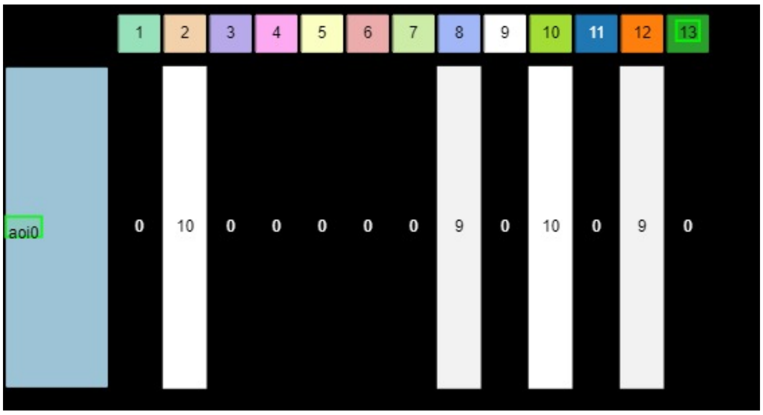


Figure 5.5: Image and AOI for the sample of Design 4



	1	2	3	4	5	6	7	8	9	10	11	12	13
aoi0	0	10	0	0	0	0	0	9	0	10	0	9	0

Figure 5.6: Matrix showing the percentage of fixations that hit the AOI for the sample of Design 4

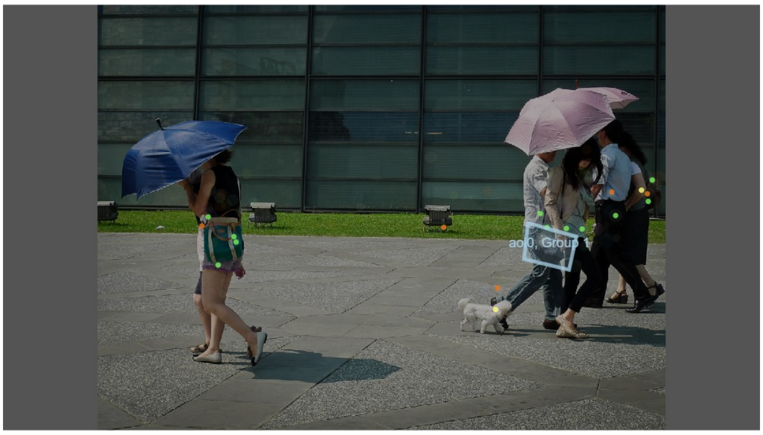
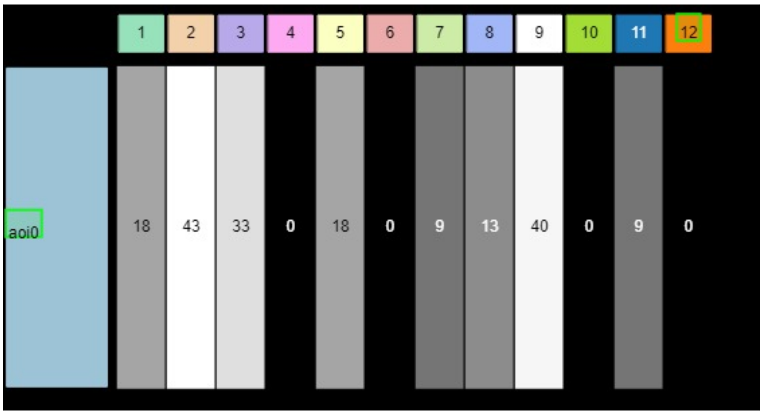


Figure 5.7: Image and AOI for the sample of Design 5



	1	2	3	4	5	6	7	8	9	10	11	12
aoi0	18	43	33	0	18	0	9	13	40	0	9	0

Figure 5.8: Matrix showing the percentage of fixations that hit the AOI for the sample of Design 5

6 Conclusion

6.1 Conclusion

This thesis explored different alternatives of study designs for eye tracking studies, with the aim of optimizing the experimental design for CL and reporting the recommendations to the research community. The approach was to alter the order of image and question, as well as the medium with which the question was presented, and conduct an eye tracking study with VQA as a task. Five different designs were created and used for comparison. The dependent variables chosen for measurement were a modification of the NASA-TLX, accuracy and the gaze data. The modified NASA-TLX was chosen as our main dependable variable, and accuracy was used as a second, indirect measure of CL. Gaze data was used for an exploratory data analysis. The study was completed by 13 participants.

The research questions were:

1. Does the presentation order of image and question significantly impact CL in an eye tracking study?
2. Does presenting questions auditory, rather than displaying them as text, affect CL in an eye tracking study?

The results of the study suggest that there is a significant impact on the CL caused by the presentation order image and question. The change of medium on the other hand did not show a significant impact on the CL.

Although the change of medium did not show a significant impact on the CL, the results still showed a slightly lower CL and a slightly higher accuracy, which is why the recommendation would be to use a textual presentation of the question and making sure the question is available to participants before the image is shown.

6.2 Limitations

The study in this thesis tested conditions within subjects, as all participants tested all conditions. With a limited amount of participants, it is justifiably a good choice, as it allows collecting a good amount of data. The disadvantage of this is that the same stimuli could not be used for all conditions, as participants would be able to memorize images and questions, which would make the designs in the later parts of the study much easier and hence would reduce cognitive load.

6.3 Future work

This thesis aim was to report recommendations for non-intrusive study design in the field of eye tracking. While some insights were gained, there is more research needed in this field, as this study only focused on the order and medium of the task description, with many design options staying consistent throughout the study.

The domains in which eye tracking is used can vary in their to be examined stimuli. In this study, stimuli were given through the underlying GQA Dataset, which consist of images exclusively. Additional research can be done on other stimuli, like paragraphs of text or real life scenes. Results might change depending on the type of stimuli.

The eye tracker used for the study might also have an effect on participants. Research can be done, to make suggestions for a specific type of eye tracker or a manufacturer to minimize cognitive load.

As more research on the topic of cognitive load theory gets published, measurements of cognitive load will become more reliable and descriptive.

A study with more participants would allow for conditions to be tested between subjects, meaning one participant would only test one condition. This would be advantages as the study would be able to use the same stimuli for different conditions, while still being able to collect a good amount of data thanks to the higher number of participants. Using the same stimuli for every condition allows for a better analysis using gaze patterns, which could give valuable insights.

I hope that the insights this thesis gave will prove useful for future research in the field of eye tracking.

Acknowledgments

First and foremost, I would like to thank my supervisors, Dr. Sandeep Vidyapu and Sita Vriend, for guiding me through all the steps to finishing my thesis, and to the 13 participants, for giving up their time to participate in my study. I would also like to thank Prof. Dr. Daniel Weiskopf for examining my work. Lastly, special thanks to Kun-Ting Chen for providing me with access to WebVeta, a tool for analysis of eye tracking data, and helping me with it along the way.

Bibliography

[20] *TLX @ NASA Ames - Home*. Dec. 2020. URL: https://humansystems.arc.nasa.gov/groups/tlx/ (cit. on p. 26).

[AAL+15] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433 (cit. on p. 13).

[AF13] A. Al-Rahayfeh, M. Faezipour. "Eye tracking and head movement detection: A state-of-art survey". In: *IEEE journal of translational engineering in health and medicine* 1 (2013), pp. 2100212–2100212 (cit. on p. 13).

[AS90] P. Ayres, J. Sweller. "Locus of difficulty in multistage mathematics problems". In: *The American Journal of Psychology* (1990), pp. 167–193 (cit. on p. 27).

[Ayr01] P. L. Ayres. "Systematic mathematical errors and cognitive load". In: *Contemporary Educational Psychology* 26.2 (2001), pp. 227–248 (cit. on p. 27).

[Bro12] J. L. Brooks. "Counterbalancing for serial order carryover effects in experimental condition orders." In: *Psychological methods* 17.4 (2012), p. 600 (cit. on p. 24).

[BWGT10] A. Bulling, J. A. Ward, H. Gellersen, G. Tröster. "Eye movement analysis for activity recognition using electrooculography". In: *IEEE transactions on pattern analysis and machine intelligence* 33.4 (2010), pp. 741–753 (cit. on p. 13).

[CJYZ20] S. Chen, M. Jiang, J. Yang, Q. Zhao. "AiR: Attention with Reasoning Capability". In: *CoRR* abs/2007.14419 (2020). arXiv: 2007.14419. URL: https://arxiv.org/abs/2007.14419 (cit. on p. 17).

[CPS08] C. Calvi, M. Porta, D. Sacchi. "e5Learning, an e-learning environment based on eye tracking". In: *2008 Eighth IEEE International Conference on Advanced Learning Technologies*. IEEE. 2008, pp. 376–380 (cit. on p. 13).

[DB14] V. DePuy, V. W. Berger. "Counterbalancing". In: *Wiley StatsRef: Statistics Reference Online* (2014) (cit. on p. 24).

[DN08] A. DiDomenico, M. A. Nussbaum. "Interactive effects of physical and mental workload on subjective workload assessment". In: *International journal of industrial ergonomics* 38.11-12 (2008), pp. 977–983 (cit. on p. 26).

[GG03] J. A. Gliem, R. R. Gliem. "Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales". In: Midwest Research-to-Practice Conference in Adult, Continuing, and Community ... 2003 (cit. on p. 32).

[Har06] S. G. Hart. "Nasa-Task Load Index (NASA-TLX); 20 Years Later". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50.9 (2006), pp. 904–908. DOI: 10.1177/154193120605000909. eprint: https://doi.org/10.1177/154193120605000909. URL: https://doi.org/10.1177/154193120605000909 (cit. on p. 26).

[HM19] D. A. Hudson, C. D. Manning. “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6693–6702. DOI: 10.1109/CVPR.2019.00686 (cit. on pp. 17, 20).

[HNKH17] R. S. Hessels, D. C. Niehorster, C. Kemner, I. T. Hooge. “Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC)”. In: *Behavior research methods* 49.5 (2017), pp. 1802–1823 (cit. on p. 28).

[HP64] E. H. Hess, J. M. Polt. “Pupil Size in Relation to Mental Activity during Simple Problem-Solving”. In: *Science* 143.3611 (1964), pp. 1190–1192. DOI: 10.1126/science.143.3611.1190. eprint: https://www.science.org/doi/pdf/10.1126/science.143.3611.1190. URL: https://www.science.org/doi/abs/10.1126/science.143.3611.1190 (cit. on p. 28).

[KDN+18] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, I. Krejtz. “Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze”. In: *PLOS ONE* 13.9 (Sept. 2018), pp. 1–23. DOI: 10.1371/journal.pone.0203629. URL: https://doi.org/10.1371/journal.pone.0203629 (cit. on p. 28).

[LMT20] J. Z. Lim, J. Mountstephens, J. Teo. “Emotion recognition using eye-tracking: taxonomy, review and current challenges”. In: *Sensors* 20.8 (2020), p. 2384 (cit. on p. 13).

[MBE95] W. F. Moroney, D. W. Biers, F. T. Eggemeier. “Some measurement and methodological considerations in the application of subjective workload measurement techniques”. In: *The international journal of aviation psychology* 5.1 (1995), pp. 87–106 (cit. on p. 26).

[MDG14] B. Morrison, B. Dorn, M. Guzdial. “Measuring cognitive load in introductory CS”. In: *ICER 2014 - Proceedings of the 10th Annual International Conference on International Computing Education Research* (July 2014). DOI: 10.1145/2632320.2632348 (cit. on pp. 16, 27).

[NAN20] D. C. Niehorster, R. Andersson, M. Nyström. *Titta: A toolbox for creating Psychtoolbox and Psychopy experiments with Tobii Eye trackers - behavior research methods*. Mar. 2020. URL: https://link.springer.com/article/10.3758/s13428-020-01358-8#citeas (cit. on p. 30).

[Nyg91] T. E. Nygren. “Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload”. In: *Human factors* 33.1 (1991), pp. 17–33 (cit. on p. 26).

[SBS+22] S. R. Steinhauer, M. M. Bradley, G. J. Siegle, K. A. Roecklein, A. Dix. “Publication guidelines and recommendations for pupillary measurement in psychophysiological studies”. In: *Psychophysiology* 59.4 (2022). e14035 PsyP-2022-0069, e14035. DOI: https://doi.org/10.1111/psyp.14035. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.14035. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.14035 (cit. on p. 28).

[SPP14] N. Steinhausen, R. Prance, H. Prance. “A three sensor eye tracking system based on electrooculography”. In: *SENSORS, 2014 IEEE*. IEEE. 2014, pp. 1084–1087 (cit. on p. 13).

[SW65] S. S. Shapiro, M. B. Wilk. “An analysis of variance test for normality (complete samples)”. In: *Biometrika* 52.3/4 (1965), pp. 591–611 (cit. on p. 34).

[Swe11] J. Sweller. “Cognitive load theory”. In: *Psychology of learning and motivation*. Vol. 55. Elsevier, 2011, pp. 37–76 (cit. on p. 16).

[Swe88] J. Sweller. “Cognitive load during problem solving: Effects on learning”. In: *Cognitive science* 12.2 (1988), pp. 257–285 (cit. on p. 16).

[UJR04] G. Underwood, L. Jebbett, K. Roberts. “Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search”. In: *Quarterly Journal of Experimental Psychology Section A* 57.1 (2004), pp. 165–182 (cit. on pp. 17, 37).

[VMKH21] K. Virtanen, H. Mansikka, H. Kontio, D. Harris. “Weight watchers: NASA-TLX weights revisited”. In: *Theoretical Issues in Ergonomics Science* (2021), pp. 1–24 (cit. on p. 26).

[VS10] T. Van Gog, K. Scheiter. *Eye tracking as a tool to study and enhance multimedia learning*. 2010 (cit. on p. 13).

[Wol21] J. M. Wolfe. “Guided Search 6.0: An updated model of visual search.” In: *Psychonomic bulletin & review* 28 4 (2021), pp. 1060–1092 (cit. on p. 15).

[Woo07] R. F. Woolson. “Wilcoxon signed-rank test”. In: *Wiley encyclopedia of clinical trials* (2007), pp. 1–3 (cit. on p. 34).

[YS75] L. R. Young, D. Sheena. “Survey of eye movement recording methods”. In: *Behavior research methods & instrumentation* 7.5 (1975), pp. 397–429 (cit. on p. 13).

All links were last followed on December 19, 2022.

A Study description and consent form

Studydescription

of "Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks? " by the Visualization Research Center of the University of Stuttgart (VISUS).

At VISUS, we conduct research on the development and design of visualizations as well as their use. The study „Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks?“ Aims to investigate cognitive load in the context of eye-tracking experiments. A complete evaluation for the intended purpose is only possible after the entire procedure has been completed. Nevertheless, **you can stop the study at any** time and go. In this case, we could use the incomplete data for partial investigations.

1. General conditions for participation:

- You have normal vision or vision corrected to normal level
- You have no physical injury
- You are at least 18 years old

2. Procedure

- Preparations:
  - You will read the ***study description***, ***privacy information*** and sign the ***consent form*** to participate in the study. You can ask questions at any time.
- Main study:
  - You answer the ***pre-test survey.***
  - There will be 3 question-image-pairs for training purpose.
  - After that the eye-tracker is calibrated.
  - There are 5 different blocks with 20 question-image-pairs each, so 100 in total.
  - You will see different images and have to give ***answers*** to questions about the images while we record your ***gaze*** on the computer screen.
  - Depending on the block the order of question and image changes:
    - Question Image Question
    - Question Image
    - Image Question Image
    - Image Question
    - As audio before, during and after the Image
  - Questions will be displayed for as long as you want.
  - Images will always be shown for a fixed 3 seconds.
  - After each block there will be a ***questionnaire***.
- In total, the study takes about 35 minutes.

3. Possible risks

During the experiment no drug is administered. The interaction devices used (mouse and keyboard) are widely used in home and work environments. Remote eye trackers emit infrared light to obtain accurate images of the user's eyes. The light intensity is not high enough to feel the heat, and so low that even sunlight from outside interferes with correct imaging. The trackers are placed at a distance of approx. 60 cm. Prolonged viewing of a computer screen can cause eye strain. To date, there are no known risks or side effects associated with short-term use of computer monitors, mice, keyboards, or remote eye trackers at these distances. The researcher will be present throughout the experiment.

4. Data use and data processing

- Pseudonymous data is identified only by an ID number. They are pseudonymous because it seems as if they belong to an abstract person and cannot be traced back to you. These data consist of ***pre-test survey***, ***gaze***, ***answers*** and ***questionnaire***.
- We have no way to connect you to the pseudonymous recordings. Therefore, we can not delete them later.
- We keep the ***consent form*** in a locked cabinet as proof of your consent. We may need to provide evidence of it in the event of an audit by funding agencies or government authorities (e.g. Deutsche Forschungsgemeinschaft DFG).

5. Further use of pseudonymous data

- We will use the data for research purposes only.
- We could reuse the data for further research at the University of Stuttgart and at its external cooperation partners.
- The data can be uploaded to public archives and shared from there.
- We cannot force others to stop sharing or limit their use of the information.
- Data processing may take place outside the European Union (EU), where data protection laws may not be comparable. This may potentially limit your rights regarding the data.

6. Contact details

In case of uncertainty, you can contact:

Researcher
Amer Rama
Germany
✉ st156339@stud.uni-stuttgart.de

Supervisor
Dr. Sandeep Vidyapu
✉ sandeep.vidyapu@visus.uni-stuttgart.de

Supervisor
Sita Vriend
✉ sita.vriend@visus.uni-stuttgart.de

Administration office
Visualisierungsinstitut der Universität Stuttgart (VISUS)
Allmandring 19
70569 Stuttgart
Germany
✉ sekretariat@visus.uni-stuttgart.de
☎ +49 (0) 711 685-88600

Studydescription

On the occasion of the data collection at the study "Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks? " by the Visualization Research Center of the University of Stuttgart (VISUS).

Responsible in the sense of data protection law
Universität Stuttgart
Keplerstraße 7
70174 Stuttgart
Germany
Telefon: +49 711/685-0
E-Mail: poststelle@uni-stuttgart.de

Data Protection Officer
Universität Stuttgart
Datenschutzbeauftragter
Breitscheidstr. 2
70174 Stuttgart
Germany
Telefon: +49 711 685-83687
Fax: +49 711 685-83688
E-Mail: datenschutz@uni-stuttgart.de

Categories of data processed
In the study we process:

- Personal data ***(marked red)*** in the ***consent form***.
- Pseudonymous data ***(marked blue)*** in ***pre-test survey***, ***answers***, ***questionnaire,*** and the recorded ***gaze*** when looking at and around the computer screen during study assignments.

Purpose of data processing and consequences of failure to provide personal data

- The ***consent form*** and pseudonymous data (***pre-test survey***, ***gaze, answers***, ***questionnaires***) are necessary for the conducting of the study and scientific publications within the framework of a research project.

At VISUS, we conduct research on the development and design of visualizations as well as their use. The study „Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks?“ Aims to investigate cognitive load in the context of eye-tracking experiments. A complete evaluation for the intended purpose is only possible after the entire procedure has been completed. Nevertheless, **you can stop the study at any** time and go. In this case, we could use the incomplete data for partial investigations.

Further use of pseudonymous data:

- Use in further research projects at VISUS or in joint research projects with external partners, e.g. in the context of collaborative research centers. https://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/index.jsp
- Publish data as part of replicable, open research.

If the above consents are not given, no disadvantages will arise. However, participation in the study is then not possible.

Legal basis

1. Conducting the research project:

Art. 6 Abs. 1 lit. e in Verbindung mit Art. 6 Abs. 3 Datenschutz-Grundverordnung (DSGVO) in Verbindung mit § 13 Abs.1 Landesdatenschutzgesetz Baden-Württemberg,

6 Abs. 1 lit. c in Verbindung mit §§ 70, 75 Landeshaushaltsordnung.

2. Consent to other uses:

Art. 6 Abs. 1 lit. a DSGVO

Recipients

- ***Pseudonymous research data*** in scientific publications: global readers/users and reviewers.
- ***Pseudonymous research data*** in a repository: global users***.***

The data may also be processed in countries that do not have a level of data protection comparable to that of the EU. There may be a high risk to your rights and freedoms in certain circumstances

- The ***consent form***: Archive of the research project.

Duration of storage

- All data will be retained for 10 years after project completion.
- Research data may be transferred from the University Archives and retained there indefinitely.

Your rights

- You have the right to obtain information from the university about the data stored about you and to have incorrect stored data corrected.
- You also have the right to request deletion or restriction of processing or the right to object to processing.
- Before deleting archived data, the consent of the University Archives must be obtained. It then decides whether the data will be deleted or retained.

For this purpose, please contact the data protection officer of the University of Stuttgart by e-mail (datenschutz@uni-stuttgart.de).

- You have the right to lodge a complaint with a supervisory authority if you believe that the processing of personal data concerning you is in breach of the law.

The supervisory authority in Baden-Württemberg is the State Commissioner for Data Protection and Freedom of Information Baden-Württemberg. https://www.baden-wuerttemberg.datenschutz.de/

Consent form

For participation in the study "Eye-tracking Study: Which Experimental Design is Better Suited for Cognitive Tasks? " by the Visualization Research Center of the University of Stuttgart (VISUS).

Please read this form carefully. In case of questions, feel free to ask the present researcher.

- I have received a copy of the ***study description*** and ***privacy information***.
- I can and read and understand English.
- I have read the ***study description*** and agree with the data usage and processing.
- I have read the ***privacy information*** and agree to it.
- I agree and **participate voluntarily**. Not participating does not result in any kind of disadvantage.
- I am free to stop participating in the study at any point. There is no disadvantage to doing so.
- During the study, I can withdraw my consent to the data collection and processing. Already collected data will be deleted.
- Withdrawing the consent after completion of the study is not possible, as pseudo-anonymous data from a specific person cannot be identified in retrospect.

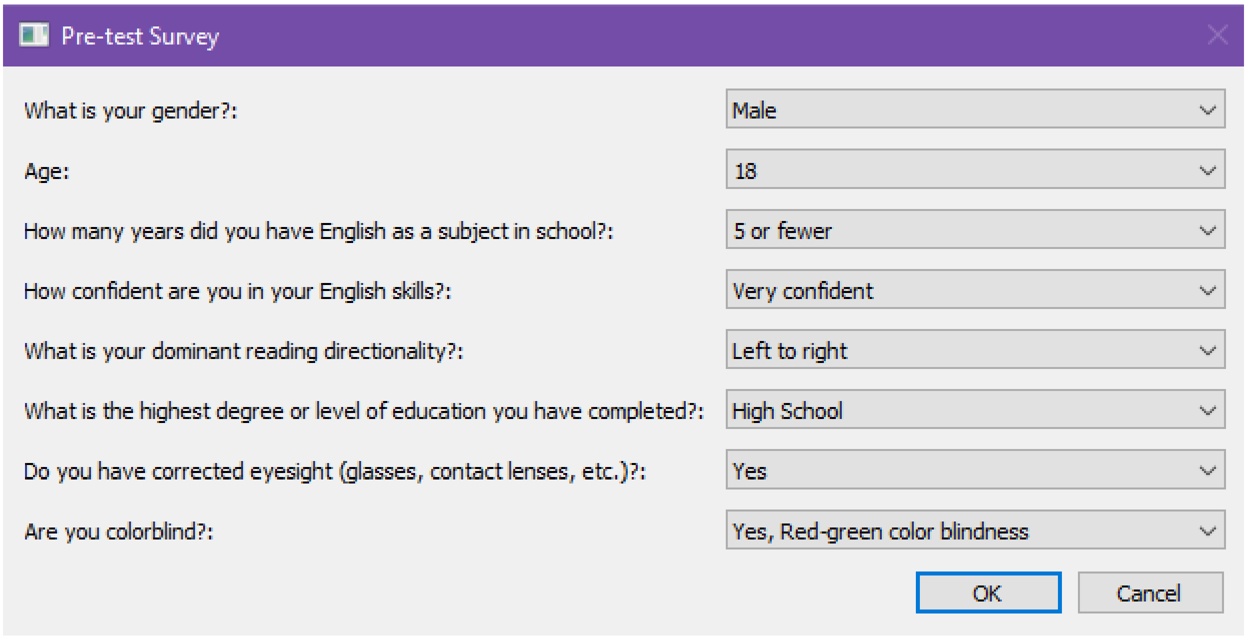
. . . .
. . . . . .

Location, date Participant name Signature

. . . .
. . . . . .

Location, date Researcher name Signature

B Pre-test survey

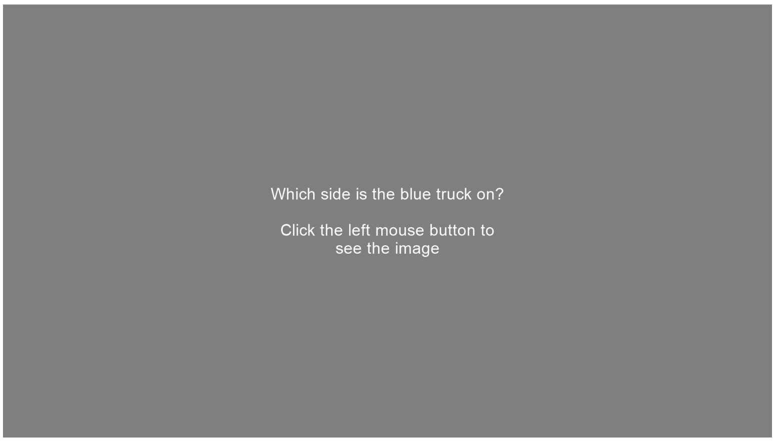


Pre-test Survey

What is your gender?:	Male
Age:	18
How many years did you have English as a subject in school?:	5 or fewer
How confident are you in your English skills?:	Very confident
What is your dominant reading directionality?:	Left to right
What is the highest degree or level of education you have completed?:	High School
Do you have corrected eyesight (glasses, contact lenses, etc.)?:	Yes
Are you colorblind?:	Yes, Red-green color blindness

Figure B.1: Pre-test survey

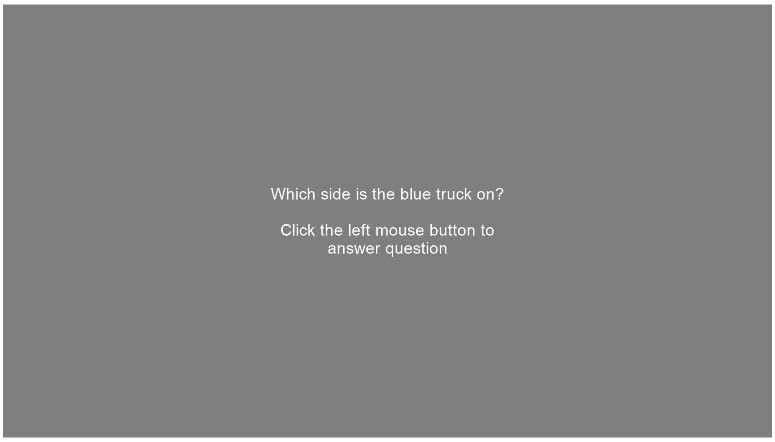
C Examples for each design



Which side is the blue truck on?

Click the left mouse button to see the image

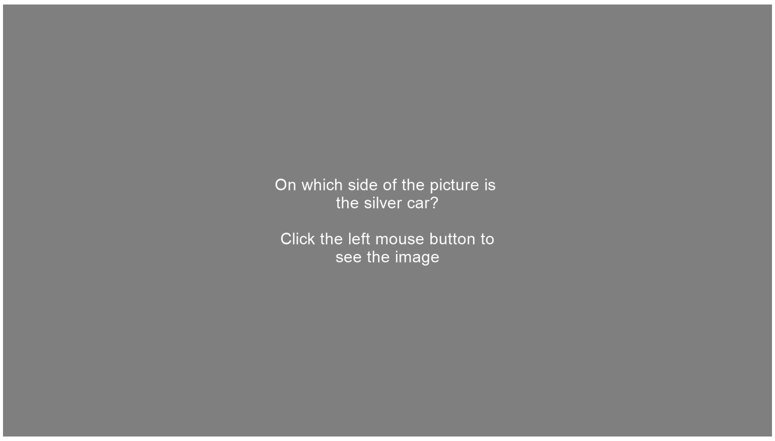
Figure C.1: Example text before image was shown for design 1



Which side is the blue truck on?

Click the left mouse button to answer question

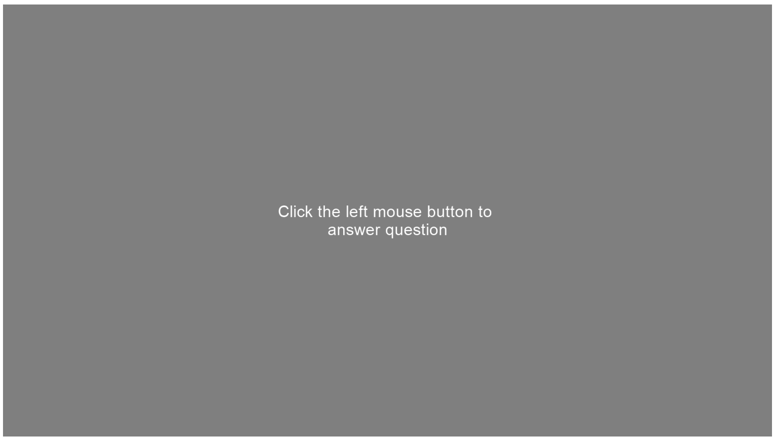
Figure C.2: Example text after the image was shown for design 1



On which side of the picture is
the silver car?

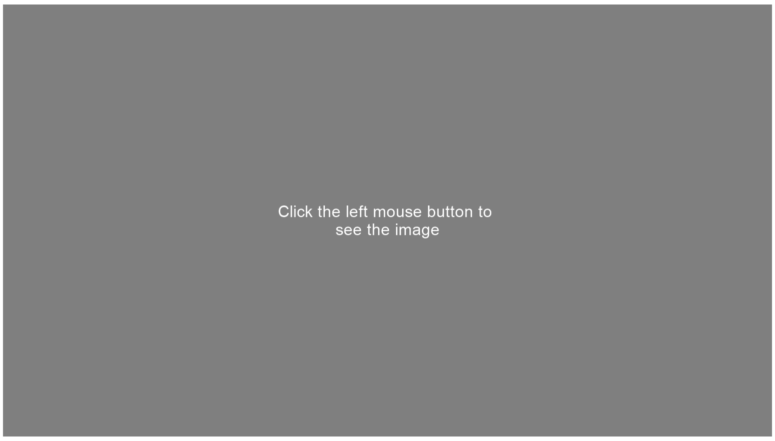
Click the left mouse button to
see the image

Figure C.3: Example text before image was shown for design 2



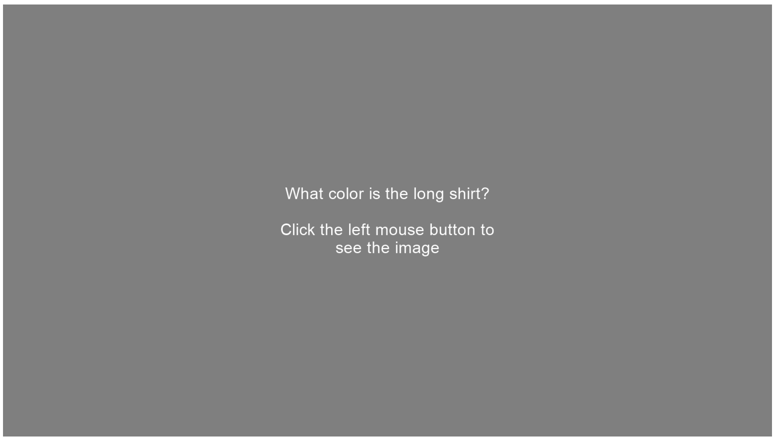
Click the left mouse button to
answer question

Figure C.4: Example text after the image was shown for design 2



Click the left mouse button to
see the image

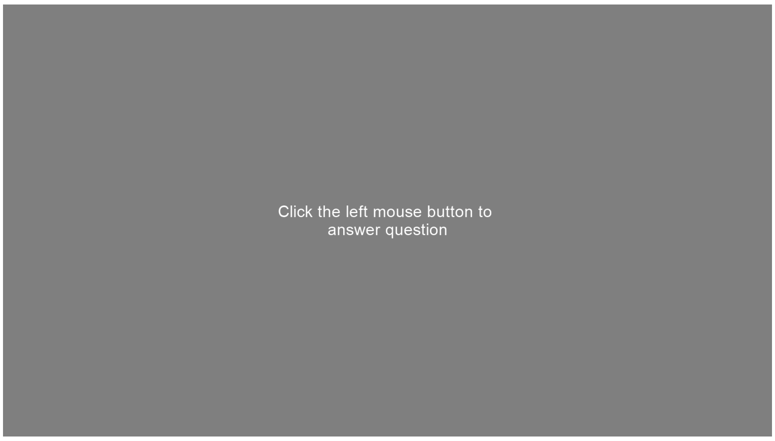
Figure C.5: Example text before the image was shown for the first time for design 3



What color is the long shirt?

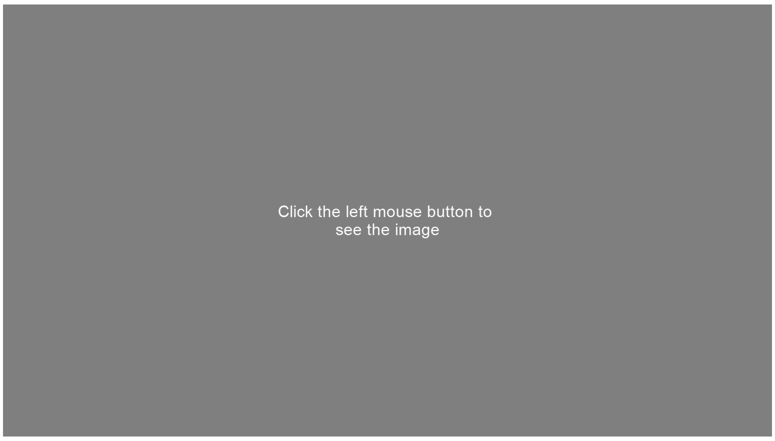
Click the left mouse button to see the image

Figure C.6: Example text after the image was shown for the first time for design 3



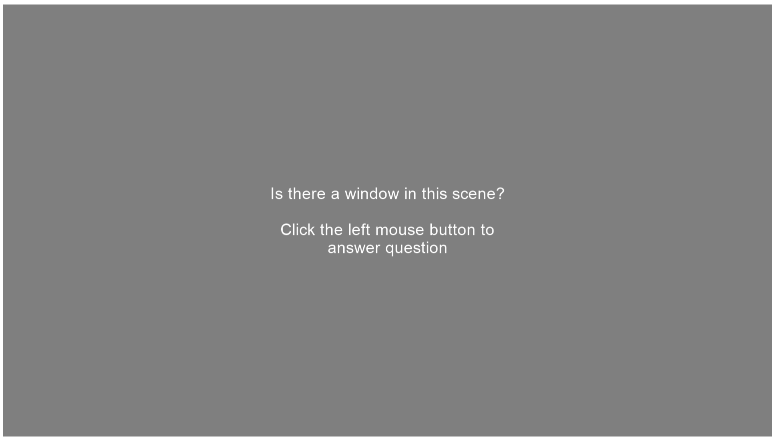
Click the left mouse button to answer question

Figure C.7: Example text after the image was shown for the second time for design 3



Click the left mouse button to see the image

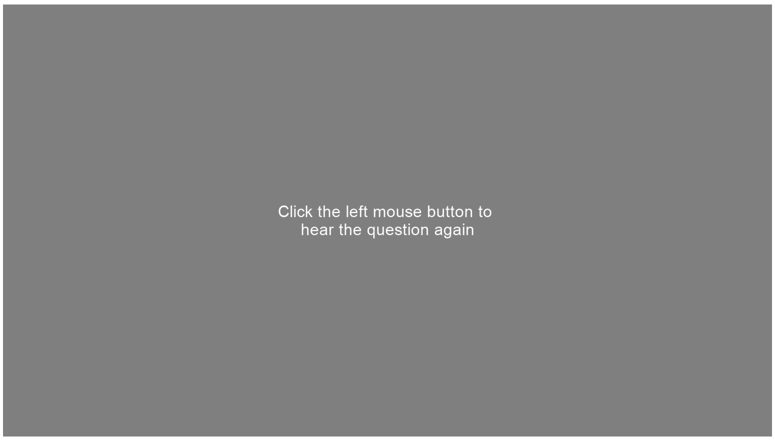
Figure C.8: Example text before the image was shown for design 4



Is there a window in this scene?

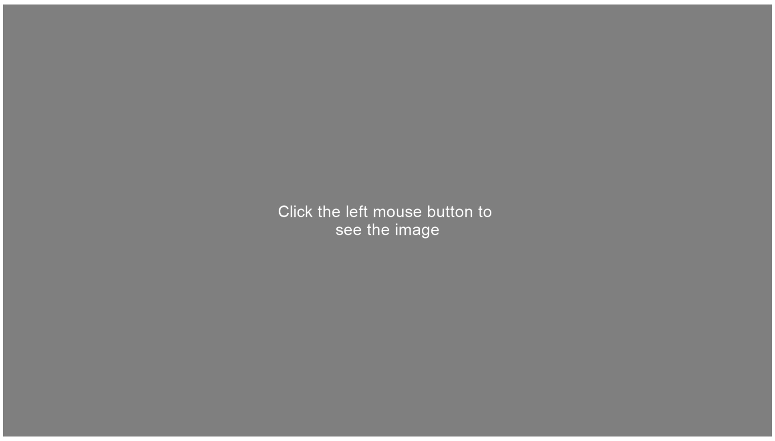
Click the left mouse button to answer question

Figure C.9: Example text after the image was shown for design 4



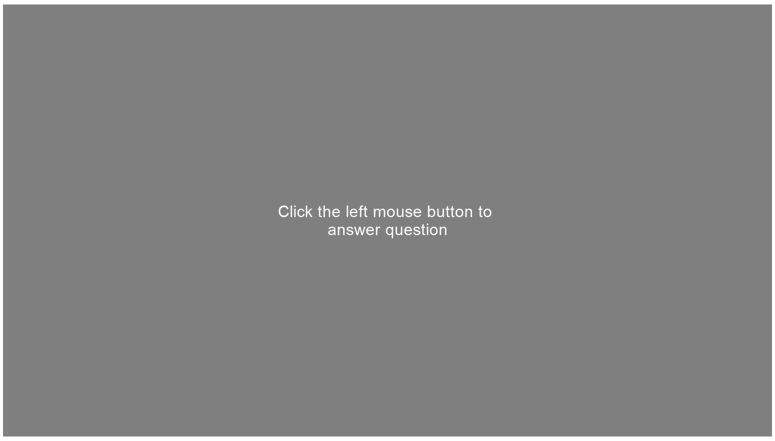
Click the left mouse button to hear the question again

Figure C.10: Example text when hearing the question for the first time for design 5



Click the left mouse button to see the image

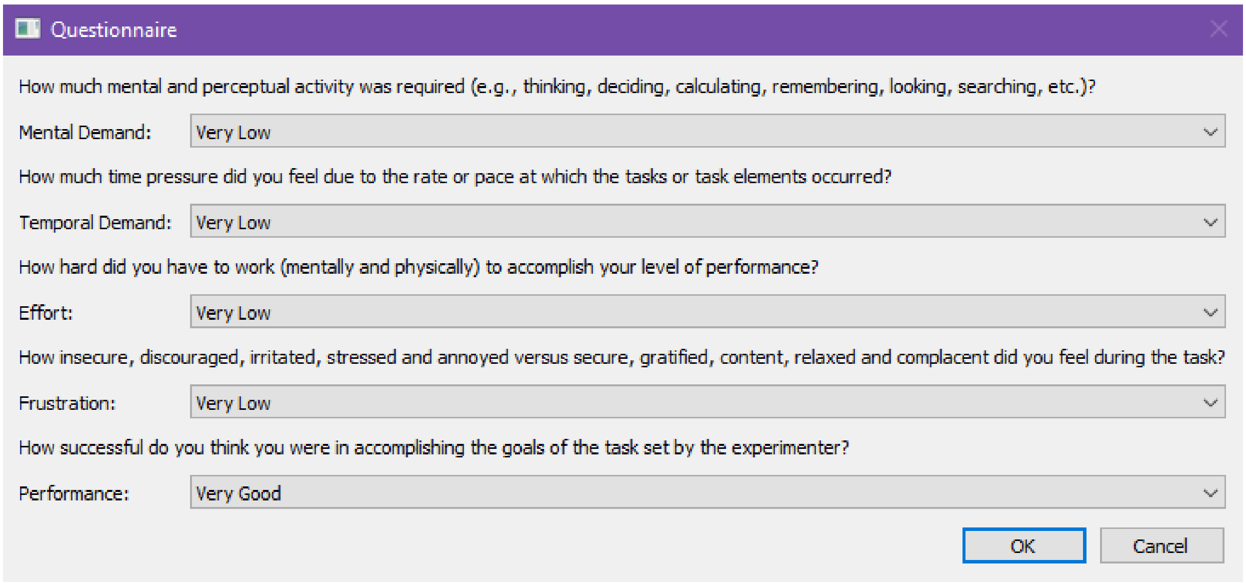
Figure C.11: Example text when hearing the question for the second time for design 5



Click the left mouse button to
answer question

Figure C.12: Example text after the image was shown for design 5

D NASA Task Load Index questionnaire



How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?

Mental Demand: Very Low

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

Temporal Demand: Very Low

How hard did you have to work (mentally and physically) to accomplish your level of performance?

Effort: Very Low

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Frustration: Very Low

How successful do you think you were in accomplishing the goals of the task set by the experimenter?

Performance: Very Good

Figure D.1: Questionnaire after every design

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

place, date, signature