stress+ - Development and Evaluation of a Web Application for Remote, Internet-based Induction of Acute Psychosocial Stress

Master's Thesis in Computer Science

submitted
by

Benjamin Zenke

born 13.10.1994 in Starnberg

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Robert Richer M. Sc., Veronika König M. Sc., Prof. Dr. Bjoern Eskofier

Started: 01.09.2021

Finished: 01.03.2022

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.
Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 01.03.2022

Übersicht

Diese Arbeit beschreibt die Weiterentwicklung der stress+ Anwendung, einer Webanwendung zur internetbasierten Auslösung von akutem psychosozialem Stress. Zu diesem Zweck wurde eine umfassende Literaturrecherche durchgeführt und die Ergebnisse in Form neuer Module implementiert und in die Anwendung hinzugefügt. Stress+ enthält nun eine Vielzahl verschiedener Stressmodule, die Designed sind um entweder psychosozialen, emotionalen oder kognitiven Stress auszulösen. Neben der Erweiterung der Anzahl der Module wurden auch mehrere Teile der Architektur der Anwendung überarbeitet. Diese Änderung betrifft insbesondere die Komponenten, die für die Ausführung und Anzeige der Stresstests erforderlich sind. Um die Ergebnisse in Bezug auf die Benutzerfreundlichkeit zu überprüfen, wurde eine abschließende Studie durchgeführt. In dieser Studie mussten die Teilnehmer verschiedene Stresstests innerhalb der Anwendung erstellen und außerdem eine Selbsteinschätzung ihrer Eindrücke abgeben. Die rekrutierten Teilnehmer hatten entweder einen psychologischen oder einen technischen Hintergrund. Obwohl die Ergebnisse Defizite in der Benutzeroberfläche aufzeigten, zeigten sie gleichzeitig eine Nachfrage nach einer solchen Anwendung, insbesondere bei den Teilnehmern aus dem Bereich der Psychologie.

Abstract

This thesis describes the further development of the stress+ application, a web application for the internet-based induction of acute psychosocial stress. For this purpose, a comprehensive literature review was conducted and the results were implemented in the form of new modules and added to the application. Stress+ now contains a variety of different stress modules that are designed to elicit either psychosocial, emotional, or cognitive stress. In addition to expanding the number of modules, several parts of the application's architecture have been revised. This modification particularly affects the components required to run and display the stress tests. In order to verify the results in terms of usability, a final study was conducted. In this study, participants had to create various stress tests within the application and also provide a self-assessment of their impressions. The recruited participants had either a psychological or a technical background. Although the results showed deficits in the user interface, they also showed a demand for such an application, especially among the participants from the field of psychology.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Laboratory Stress Testing 5

2.2 Virtual and Remote Stress Testing 7

3 Methods **9**

3.1 Stress+ Web Application 9

3.1.1 Application Features 10

3.1.2 User Interface 10

3.1.3 Application Architecture 13

3.1.4 Modifications of the Architecture 17

3.1.5 Screen and Overlay Modules 19

3.2 Evaluation 24

3.2.1 Study Design 24

4 Results **27**

4.1 Task Execution 27

4.2 Questionnaires 32

4.2.1 System Usability Scale 32

4.2.2 NASA TLX 33

4.2.3 AttrakDiff 35

4.2.4 User Experience Questionnaire 35

5 Discussion **37**

6 Conclusion and Future Work **41**

List of Figures **43**

List of Tables **45**

Bibliography **47**

A Acronyms **53**

B Task Instructions **55**

Chapter 1

Introduction

In today's world, we all face stressful situations in our daily lives. Stressors occur almost everywhere as we navigate our complex society with diverse interests and expectations. An upcoming deadline, an important presentation at work, or interpersonal stressors such as marital problems, arguments within the family, or caring for a family member are stressors we are familiar with. However, they can also be less obvious, as any subjectively challenging experience can be a stressor, whether it is positive or negative. For example, living in an area with a high crime rate can also be stressful for an individual. Nevertheless, all these situations have in common is that they require short-term adaptation. Fortunately, the human body is prepared and well-equipped to deal with them.

Acute stress activates two response pathways in our bodies [Dic04]. The results can be measured in various physical markers. The first response pathway is the sympathetic branch of the Autonomic Nervous System (ANS). It is responsible for increased heart rate, blood sugar levels, blood pressure, and inflammatory cytokines. These reactions occur rapidly and were first described as fight-or-flight reactions by Walter Cannon, as they are responsible for the increase in performance we experience when being stressed [Sap00]. These changes are then regulated and adjusted via the second pathway, the so-called Hypothalamic-Pituitary-Adrenal (HPA). The HPA axis releases a cascade of hormones, beginning with the corticotropin-releasing hormone (CRH) from the hypothalamus, adrenocorticotropic hormone (ACTH) from the pituitary gland, and cortisol from the adrenal cortex [Ded05]. Adjustments by the HPA axis are, for example, to prolong the increases in blood pressure and glucose, but also inhibiting the activation of some immune mechanisms originaly activated by the ANS [Sap00]. Compared to the ANS pathway, the HPA axis reacts much more slowly, and it can take up to 20 minutes for a response to peak. Research has shown that psychosocial stress is one of the strongest types of stress to activate the HPA axis [Kir94; Dic04].

Although the stress response serves to survive or cope with the situation, stress can also damage the organism [McE98]. Chronic stress, in particular, has been shown to be a factor in the cause of several different illnesses, such as depression and mental disorders, and life-threatening diseases, such as cardiovascular disease, insulin insensitivity, and cancer [Coh07]. And as our modern lives become more hectic and stressful, this problem is growing in magnitude, especially in industrialised countries. For example, the economic damage caused by stress is enormous: In the United States, an estimated 42 billion dollars are already spent on absenteeism and the treatment of stress-related diseases [Kal02].

However, extensive research is still needed to understand the relationship between the above stressors, the described reaction and processing of acute stress, and chronic stress development [Roh19]. Rohleder theorised that repeated exposure to acute stressors and the associated recurrent stimulation of the HPA axis and high cortisol levels over a prolonged period of time lead to chronic stress. Therefore, he proposed researchers must repeatedly induce acute stress on subjects and observe the changes and responses to understand this transition better. Unfortunately, current stress protocols in the lab are not suitable for replication due to the high requirements, such as standardised rooms and high personnel costs. Furthermore, in the past two years in the pandemic, tests in the laboratory were hampered by contact restrictions. This also interfered with further research in this area and increased the need for alternative psychological stress protocols.

For this reason, the goal of this thesis is to develop a platform for the creation of internet-based stress tests that allows to conduct studies without requiring participants to be repeatedly invited to a laboratory. To do this, this work builds on the existing framework of the “stress+” web application for acute stress induction, which was initially developed in preceding student projects. For this purpose, the existing code base was first improved and extended by additional features. As next step, new stressor modules were identified and implemented. The developed application was then evaluated in a user study to validate its general usability and how easy it is for users with different backgrounds (technical or psychological) to create their own stress pipelines. Ultimately, we hope that approaches such as this can provide an alternative to classic laboratory stress protocols such as the Trier Social Stress Test (TSST) or the Montreal Imaging Stress Task (MIST) in the future.

Chapter 2

Related Work

The preceding literature review revealed that previous research has discovered and used a variety of stressors. In psychology, a stressor is an event or environment that a person perceives as demanding, challenging or a threat to one's safety [Dec18]. A good entry point for an overview are meta-analyses of scientific stress protocols [Dic04; Noa19]. Figure 2.1 provides a overview of the types of stress and provides examples of the most commonly encountered stressors. Physical stress, for example, is our body's response to pain, while cognitive stress occurs when our brain is overloaded and given a complex task that exceeds our brain's capacity. Negative affective states, into which an individual is intentionally driven, trigger emotional stress. Psychosocial stress occurs mainly in situations where we are afraid of losing something or embarrassing ourselves. Previous research has shown that different stressor categories can trigger different physiological and psychological stress reactions. For example, as mentioned in the Introduction, psychosocial and socially-evaluative stressors can trigger a strong HPA axis response [Kir94; Dic04]. In contrast, physiological stress and also cognitive stress response primarily occur in the ANS pathway. This fact plays an essential role in designing stress protocols and the responses expected or exhibited by test participants [Dic04], not only for this thesis but also for previous researchers who have done substantial work in stress testing.

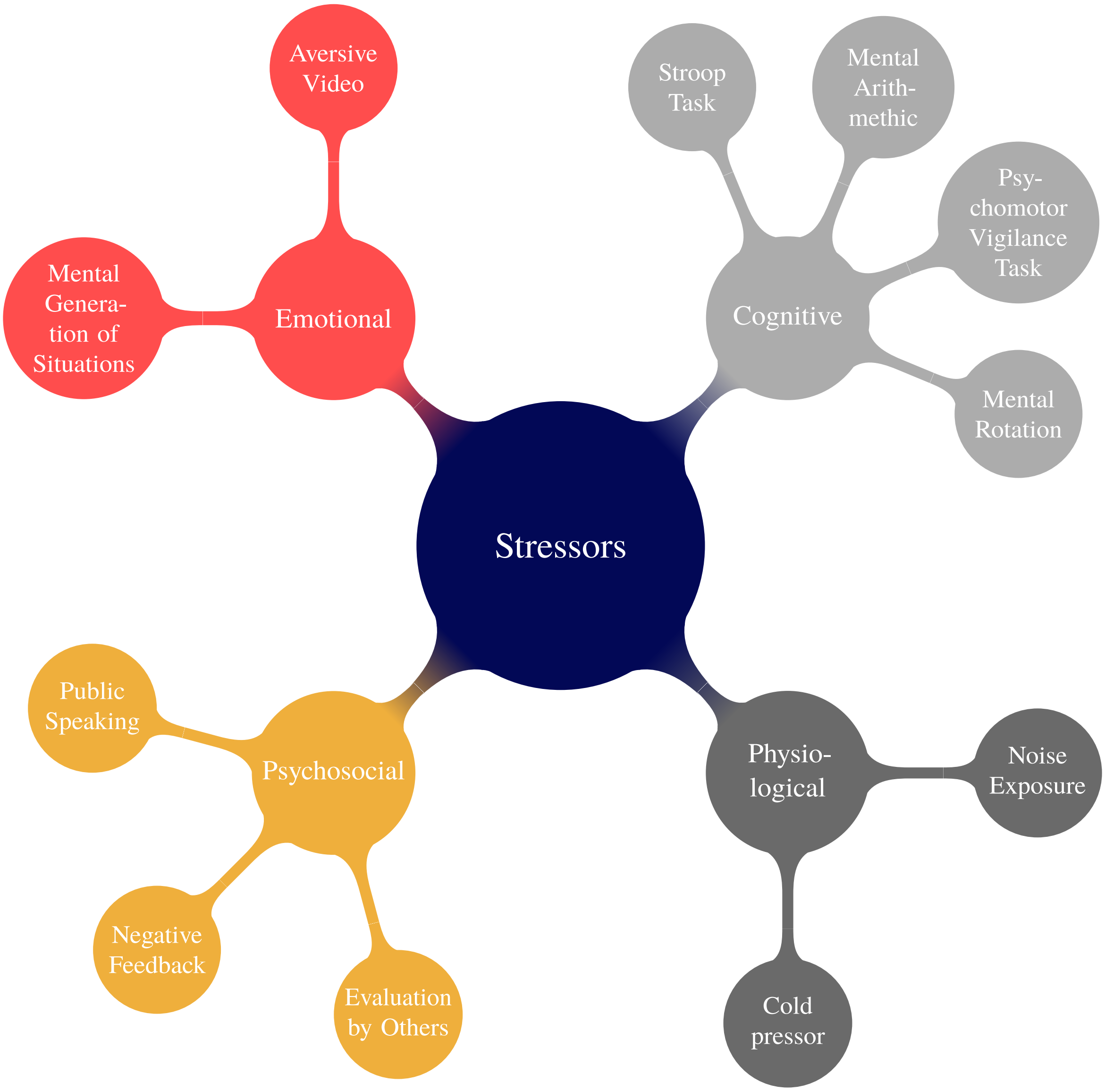


Figure 2.1: Stressors used in common stress protocols [Dic04]

Phase	**Duration** (Minutes)	**Room**	**Description**
Rest	10-30	A	Participant rests after arrival at the laboratory
Introduction	-	B	Participant stands in front of the jury, video recorder and voice analysis system - Receives the task for a job interview situation
Preparation	10	A	Time to prepare the talk - pen and paper are handed out for notes
Speech	5	B	Interview speech in front of the jury - the written notes must not be used - if the speech is too short, the probands are reminded to continue
Mental Subtraction	5	B	Participant count down from 1022 in steps of 17 as fluently as possible - At each failure they start again at 1022
Rest	30-70	A	Debriefing: Purpose of the study - no recording/voice analysis

Table 2.1: The Trier Social Stress Test [Kir93]

2.1 Laboratory Stress Testing

Stress testing is a method of exposing people to acute stress. Typically, these stress test protocols require a laboratory environment. Various physiological markers are measured during the protocol to assess the response to acute stress. These markers may include heart rate and free cortisol levels in saliva samples. Two of the most common stress tests used to assess psychosocial stress are the TSST [Kir93] and the MIST [Ded05]. Table 2.1 outlines the procedure of the TSST. This stress protocol reliably elicited physical stress responses in six independent studies prior to and many studies following its initial publication [Dic04]. On the ANS pathway, Kirschbaum et al. measured a significant increase in heart rate. As for the HPA axis, they measured significant changes in ACTH, growth hormone (GH), and free cortisol in saliva and serum samples. These results made it the gold standard protocol in stress research, and it has been reused and modified for other studies. However, it also has the disadvantage of being extremely resource-intensive. On the one hand, the TSST requires at least three researchers to conduct the testing protocol, two of whom require specialized training for the social-evaluative task. Furthermore, it requires much space since the protocol needs at least two laboratory rooms, one for the preparation and resting phase and another

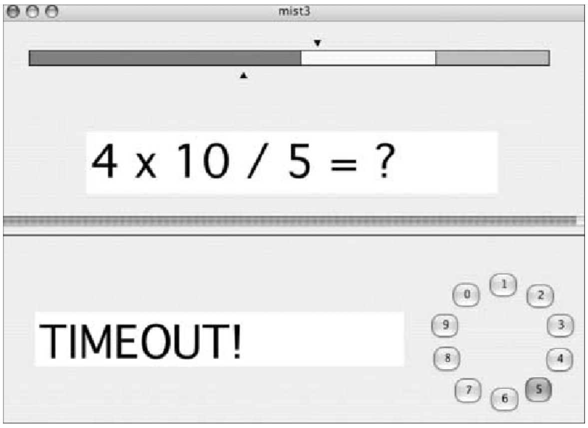


Figure 2.2: User Interface of the MIST [Ded05]

for the interview and mental arithmetic task. In addition, it is time-consuming for both the researchers and the participants. Moreover, as previous studies have shown, it cannot be used for repeated stress tests the way Rohleder [Roh19] intended to do, because a habituation effect in the HPA axis already occurs after two exposures [Pet12; Wüs05].

In 2005, Dedovic et al. [Ded05] built on the TSST and introduced the MIST. Compared to the TSST, the MIST protocol is an improvement in terms of resrouces, requiring only one research assistant, a screen, and an input device. It was originally developed for use during fMRI scans, i.e., environments where one cannot move around and the possible stressors are limited. The protocol's core is a computer screen that presents various mental arithmetic tasks. The participants must then press the correct keys on the input device within a time limit. Figure 2.2 shows a typical MIST screen. The tasks always have a result between zero and nine. The key to social-evaluative stress induction is the combination of mental arithmetic with social assessment components. During the stress test, participants already receive negative feedback by comparing their performance to the average performance of all participants. The displayed average performance is always better than the participant's current performance. Usually, a protocol run consists of several iterations with mental arithmetic tasks. During breaks, a research assistant enters the scanner room and asks the subjects to improve their performance, since otherwise their results will be useless and expensive time in the scanner is wasted. In addition, the research assistant tells the participants that they will be observed and evaluated by a board of judges. Dedovic et al. reported that they collected saliva samples and found significant increases in cortisol levels in all three studies conducted. Although the increases were smaller than after the TSST, this shows that the constant feedback from the computer program combined with the intermittent feedback from the investigator can induce acute psychosocial stress. This finding is critical to this thesis. Although the MIST requires fewer resources than the TSST, it still requires at least one research assistant, a room for testing, and participants still need to come to the lab to take the test. Consequently, the MIST cannot be used to administer multiple tests simultaneously without a great deal of additional effort. In addition, as far as it is known, there is no research on the habituation effect for the original MIST. However, another study shows a similar effect for a modified version as for the TSST [De 21].

2.2 Virtual and Remote Stress Testing

To solve the problems mentioned above with existing laboratory test protocols, previous research has attempted to adapt various approaches, most of which already exist, to a virtual or remote environment. The literature review revealed that most approaches adapted the TSST protocol.

For example, Zimmer et al. [Zim19] used a head-mounted display to present the TSST. For this, they recreated the laboratory environment and modelled it precisely in a Virtual Reality (VR) application. Similar studies on the adaptation of the TSST to VR, but often using other VR platforms, have been conducted by various research groups [Kel07; Shi16; Rui10]. Interestingly, most of them, conducted before Zimmer et al.'s research, could not elicit a robust stress response. Zimmer et al., however, were able to elicit a measurable response using saliva measurements of cortisol and alpha-amylase comparable to an in-vivo TSST. They suggested that this was probably due to improved technology, advanced graphics, increased details and realism, and improved sense of presence in their version. Of course, there are advantages to this approach, as it allows for maximum experimental control and improved standardisation. Furthermore, this approach reduces the required resources. For instance, the minimum number of experimental assistants is reduced from three to one, who also does not need to be specially trained. Furthermore, this stress protocol can also be carried out in a single room. Nevertheless, it does not solve all the problems and even creates new ones. Participants still need to be present in the lab. Otherwise, additional logistics and instruction are required for the VR displays because head-mounted displays are still specialised hardware, which is not available everywhere. Additionally, the criteria of repeatability must be kept in mind.

Therefore low habituation is required. Jönsson et al. [Jön10] found that their virtual version of the TSST was still affected by a habituation effect of the HPA axis and, consequently, a reduced cortisol response. They were able to demonstrate habituation after only two sessions.

Eagle et al. [Eag21] proposed another idea for modifying the TSST. Their approach was to administer the TSST via an online video conferencing tool. This study was conducted in response to contact restrictions during the COVID-19 pandemic and was a simple solution to avoid participants having to come to the lab. In this experiment, they were able to elicit a reliable stress response by measuring the change in heart rate during the test and self-reports after the stress test. In order to investigate habituation, they repeated the test after eight weeks. They found habituation in self-reported stress but no evidence of habituation in cardiovascular reactivity. This result is consistent with findings from other studies examining repeated measures of stress protocols [Sch03; Kän06]. It is known that habituation occurs mainly in the HPA axis. However, they also mention in their study that they only measured heart rate and did not assess the HPA axis response. Furthermore, this version also has advantages and disadvantages. The most lucid advantage is that people do not have to be present in the lab. But unfortunately, it requires the same number and training of research assistants as the original laboratory TSST and is, again, quite time-consuming. They also reported problems with the stability of the internet connection and the video conferencing tool in some runs.

The examples given here show that many researchers working on the development of reliable virtual or remote stress tests are trying to use modern technologies. However, to strive for improvement, this research will not address another TSST adaption because, as the literature review shows, the number of attempts to move away from the existing, standardised and well-proven laboratory test protocols is small. So this thesis will attempt to fill this gap by proposing a tool that might eventually allow researchers to perform stress tests remotely. Therefore, the aim of the application is to be easy to use, to elicit a reliable stress response, and for the tests to be repeatable with little effort and hopefully little habituation to the different stressors.

Chapter 3

Methods

This work builds on an existing code base and develops it further by adding new modules and making some behind-the-scenes improvements to the architecture and user interface. The following chapter describes the exact methods.

3.1 Stress+ Web Application

The Stress+ web application was initially developed as part of the ”Innovation Lab for Wearable and Ubiquitous Computing” course at the Machine Learning and Data Analytics (MaD) Lab by a group of students in the summer of 2020 in collaboration with the Chair of Health Psychology at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). This first version already pursued similar goals as this research but was limited to mapping the MIST protocol. Furthermore, the user interface was very simplified and inconsistent. At this point, Pipelines, Screens and Overlays have already formed the application's core. A Pipeline is a series of Screens or Overlay modules that describe the flow of a stress test protocol in the tool. Each stress test has one Pipeline for Screens and one for Overlays. Screens are usually responsible for displaying tasks the user needs to perform, such as the MIST math screen. On the other hand, Overlays are additional information or elements on the monitor. For example, these can display different types of feedback. This basic version was then improved by Nina Mürschberger during her computer science project. This basic version was then further extended by Nina Mürschberger during her computer science project. She added authentication and user management features as well as improvements to the user interface.

3.1.1 Application Features

The essential function of the app is the creation and execution of stress tests. For this purpose, researchers can create so-called Pipelines according to the scheme mentioned above. When creating them, they also can individually configure each module within a Pipeline via a corresponding configuration menu. These settings can include, for example, display durations, difficulty levels, but also individual texts. Once the researchers have created the protocol, they can send a link to the participants in their study, who can then perform the stress test. Nevertheless, in addition to creating and running pipelines for stress protocols, the application offers even more features. A login and authentication function has been implemented to protect the stress tests from unauthorised access as an additional feature for researchers. This user management also ensures that each researcher only has access to their tests and results. However, if researchers want to grant access to other users registered with stress+, they can use the ”Share” function to share a protocol. The other researcher will then have direct access to the settings and all results collected up to that point via their account. In addition, each test administrator also can download the recorded participants' results separately for later analysis. Furthermore, it is possible to export saved stress tests and import new stress tests into the tool via a configuration file. These configuration files contain JSON objects in which the name of the test, all added modules and their settings are stored. However, they do not contain any test results that have already been collected. Instead, this feature provides an efficient way to store the exact configuration of a particular test pipeline for archiving and potential reproducibility by other researchers.

3.1.2 User Interface

The user interface of the web application has two access points. The first access point is a login area for researchers to create and manage their stress protocols, while the seoncds access point allows study participants access via the link to a specific stress test to perform a stress test run.

After logging in, researchers land on the *Management Page*. Here, they see a list of all test protocols they have created or other users have shared with them. Researchers have the options to edit, copy and delete their test protocols. As soon as they want to create or edit a test, researchers will be forwarded to the *Editor*. Figure 3.1 shows a screenshot of the editor page to better illustrate the following section. In the upper half of the screen is a selection

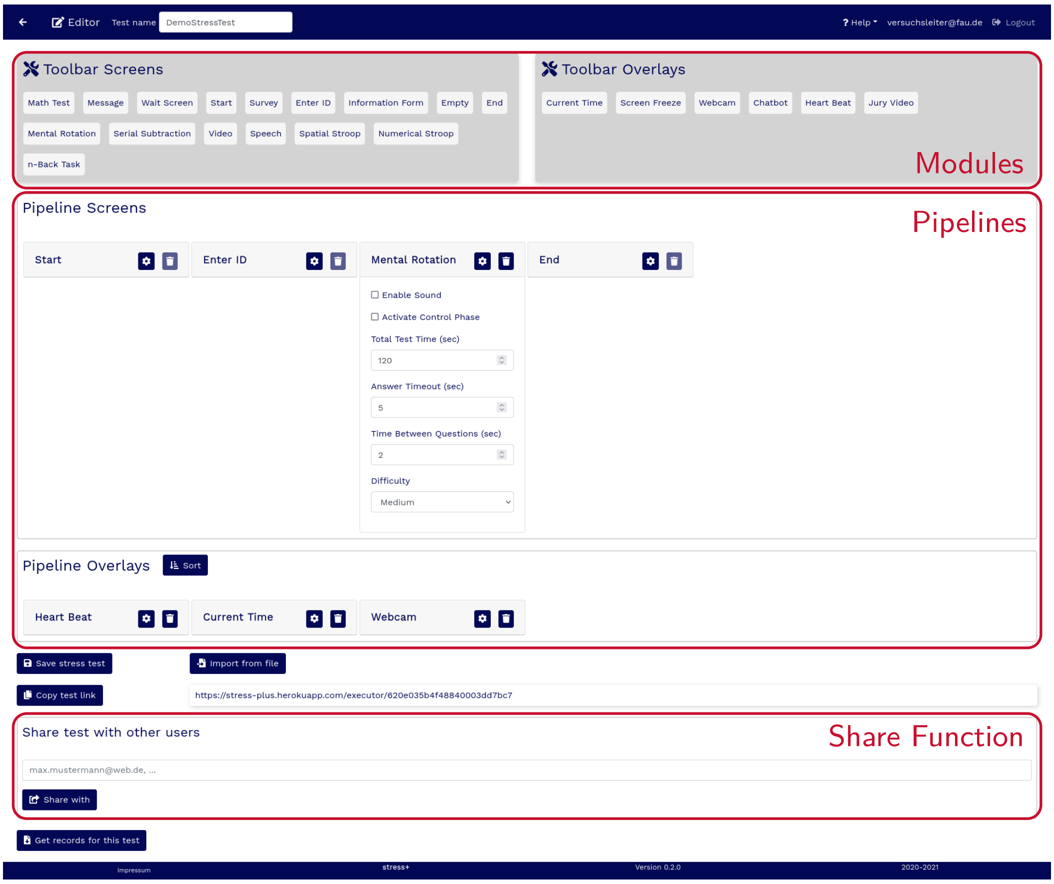


Figure 3.1: User interface for researchers

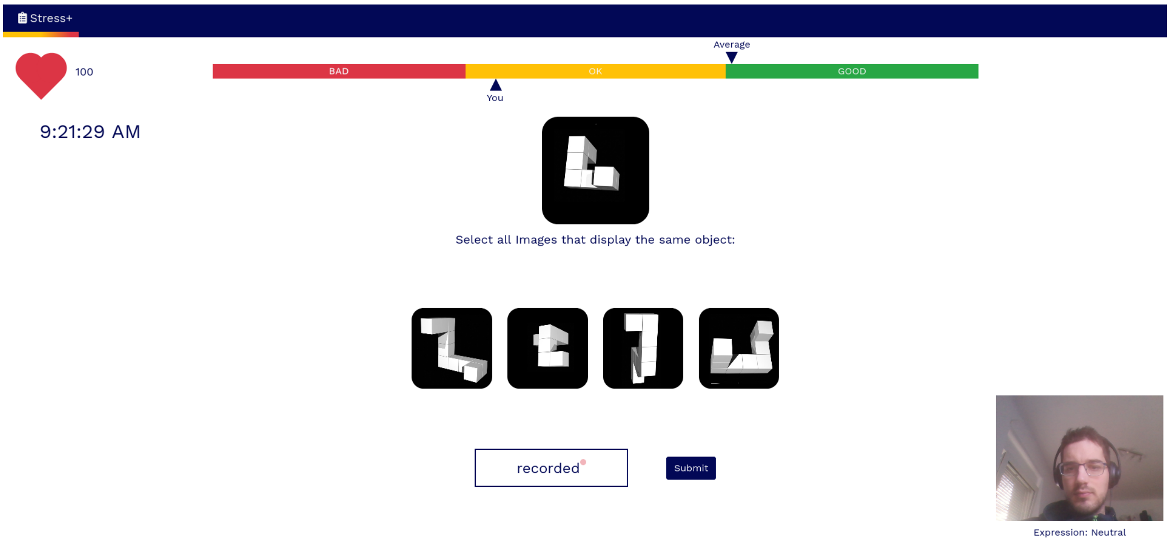


Figure 3.2: User interface for Participants showing a *Mental Rotation Task* and a *Webcam*, a *Heartbeat*, and a *Current Time Overlay*

of all the screen and overlay modules available in the application. These can be added to a protocol by dragging and dropping an element into the appropriate pipeline. Each pipeline can contain an unlimited number of modules. In addition, each module has an options menu that unfolds by clicking on the settings wheel. Below the Pipelines are the buttons for saving the protocol and importing a configuration file. As soon as the protocol is saved for the first time, the latter switches from import to export. These buttons are followed by the input field for the release function. Here, additional researchers can be granted access to the protocol, and all others who already have access are displayed. Last but not least, below the area for the share function, there is the button "Get records for this test" to download the test results.

The participants' entry point into the user interface is the so-called *Executor*. They access it via a link they receive from the test administrators. The *Executor* is responsible for running the test and the correct display of the pipelines. Each stress test starts with a welcome screen, followed by screens and overlays configured by the researchers. After that, it depends on what the creator of each pipeline has used. Each screen can also contain overlays displayed at the edge of the screen. The number of overlays can vary from screen to screen. Figure 3.2 shows an example of a mental rotation task together with overlays for the heartbeat, current time and webcam.

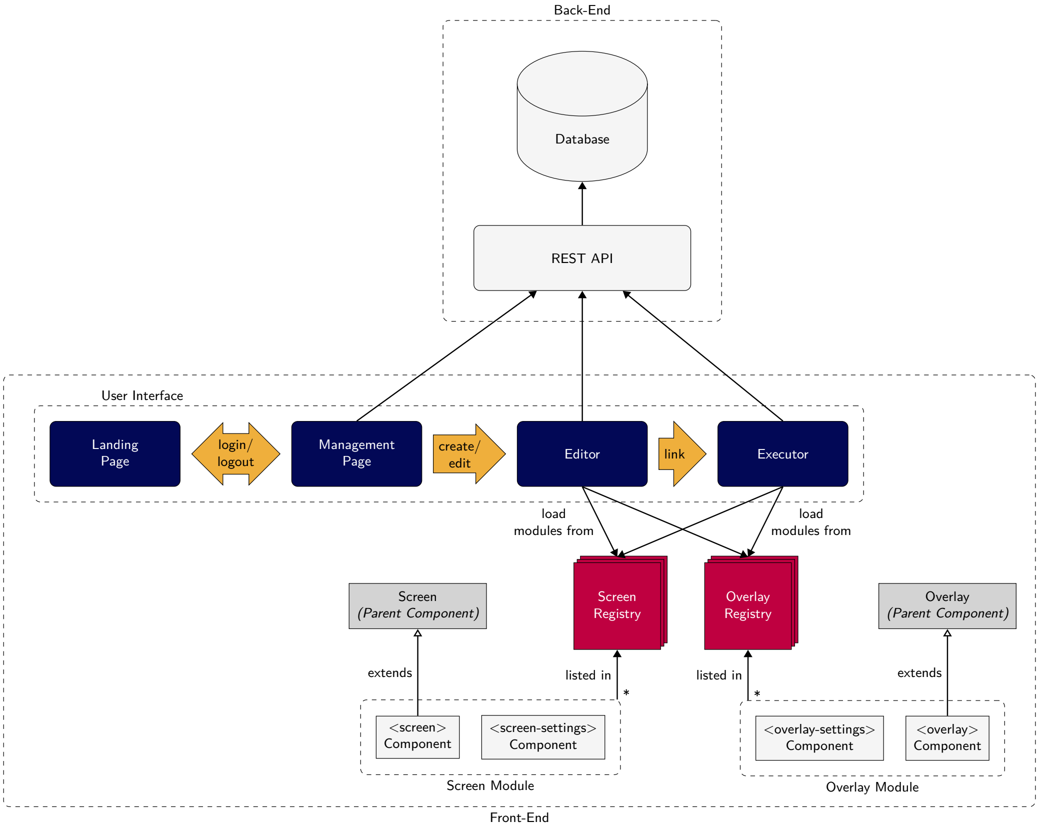


Figure 3.3: Application Architecture

3.1.3 Application Architecture

The application consists of two main parts, a front-end application and a back-end application. Figure 3.3 shows the architecture of the application, including all important components of the system.

The front-end is a browser-based single-page application written in JavaScript using the React[1] framework. The former developers designed the front-end architecture generically so that knowledge of the final available screen and overlay modules is not required. The essential components to achieving this are the *Screen* and *Overlay Registries*. To retrieve the currently available modules, both the editor and the executor query the registries and load the components accordingly. The registries hold the actual module component and an associated settings component along with the name and default settings for that module. The *Screen*

[1] https://reactjs.org/

Registry additionally stores whether a Screen is mandatory, uses the progress bar or allows Overlays. The *Overlay Registry* keeps whether an Overlay is displayed outside the defined containers. For example, this could be an overlay that requires editing the entire display area. This concept makes it straightforward to add further modules without modifying the remaining code. The only action required in this architecture is to add new modules to the corresponding registry. As described in the previous chapter, the editor is the page where the test protocols are created. Therefore, it shows all available modules from the registry and stores the created stress protocol in the backend. Stress protocols are implemented as lists for both the Screen Pipeline and for the Overlay Pipeline. Both lists include the type and configurations for each item. To keep the application generic, the module types in the list are defined by a string property containing the modules name from the registry.

As already mentioned in Section 3.1.2, the executor controls the entire process and the display of a stress test. To do this, it first loads a saved stress test from the back-end using a unique ID, which is included in the link. Thereby, it receives the screen and the corresponding overlay lists. The *Executor* then iterates over these lists to display one screen after the other. The currently displayed Screen Module and its Overlay Modules are then queried in the registry. Switching between two consecutive screens is done by replacing the current component with the subsequent one. After each screen the *Executor* also sends the collected statistics, such as given answers and time needed, back to the API. This process continues until the end of the list is reached, and the test protocol terminates. In this way, the executor is never exited or reloaded during an execution.

On the other side, the back-end is also written in Javascript and uses the Node JS platform. The front-end can access it via a REST API. The API (Figure 3.4) offers endpoints for creating, updating, loading stress tests, and saving the records during stress test execution. Data exchange between the frontend and the API is done via HTTP and uses JSON documents. The back-end stores the received data in a MongoDB[2] database. MongoDB is a document-oriented NoSQL database that allows for maximum flexibility. The system additionally uses the Javascript library Mongoose[3] to implement the structure. The database contains three collections: One to save the stress test configuration, one to save the statistic records and one to save the rule policies, i.e., the read and writes permissions of users. Figures 3.5, 3.6, and 3.7 show the structure of each table and the corresponding schemas.

[2]https://www.mongodb.com/

[3]https://mongoosejs.com/

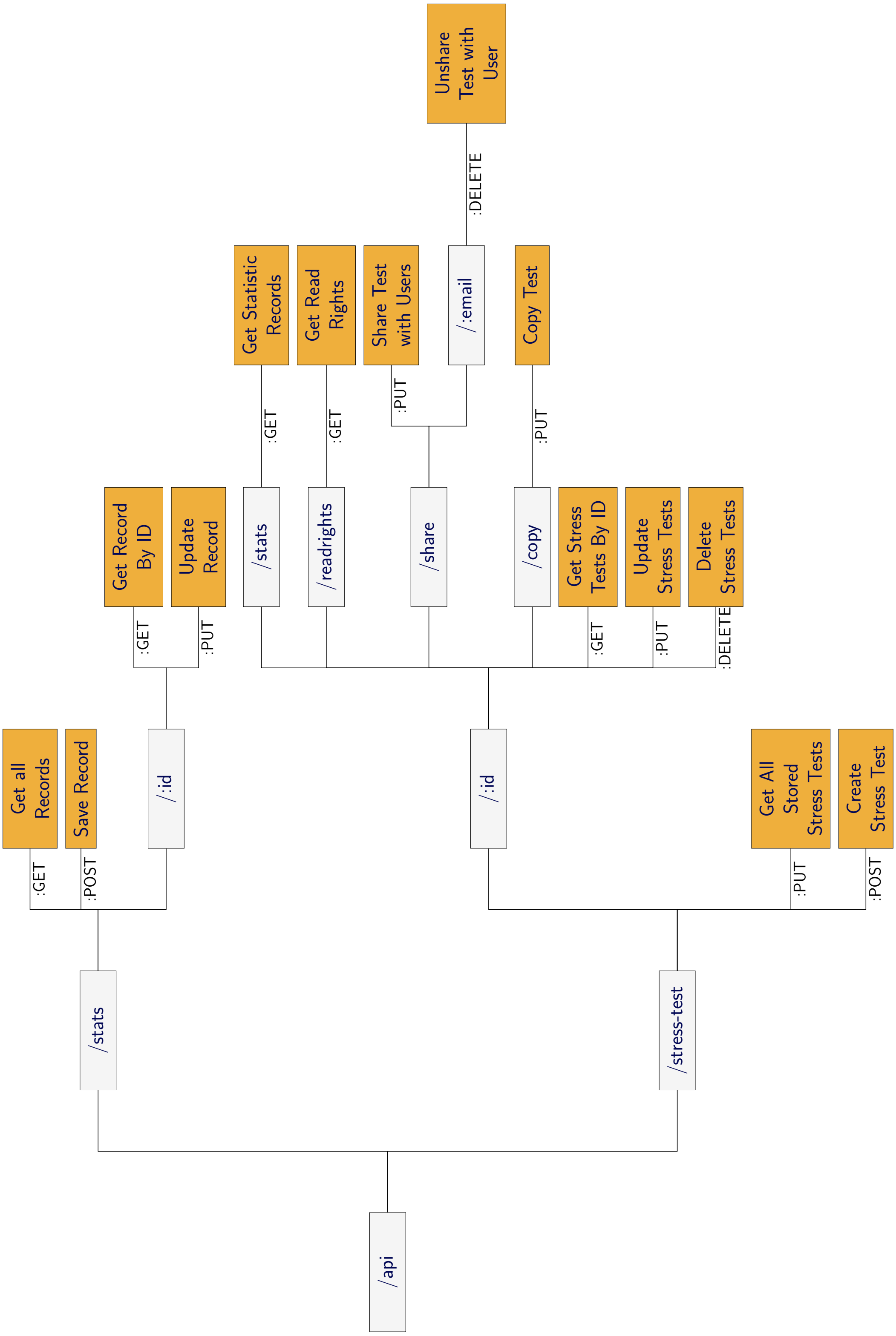


Figure 3.4: Routes of the REST API

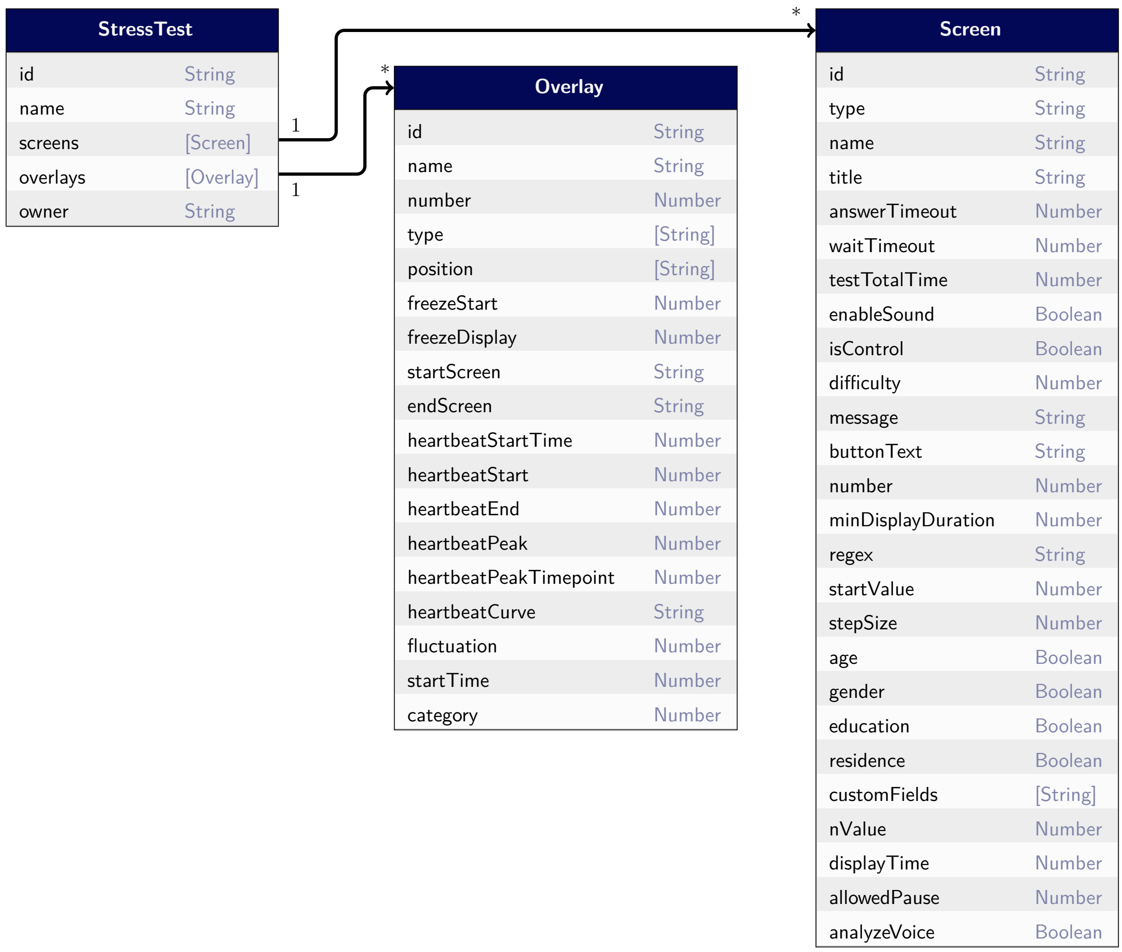


Figure 3.5: Schema of Stress Test Collection

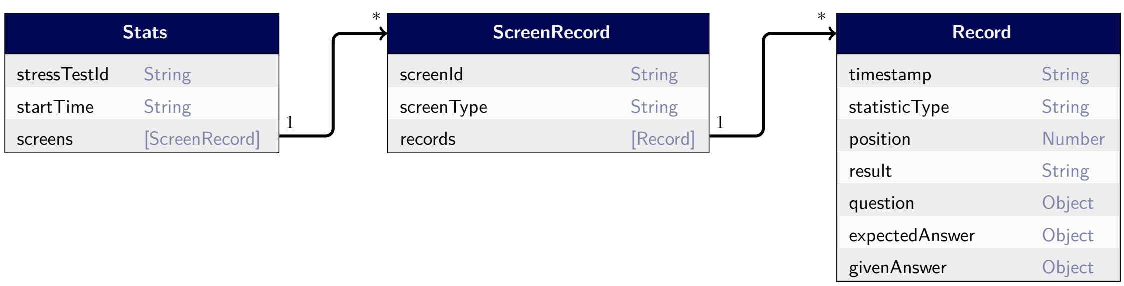


Figure 3.6: Schema of Stats Collection

RulePolicy	
userEmail	String
allowedToRead	[String]
allowedToWrite	[String]

Figure 3.7: Schema of Rule Policy Collection

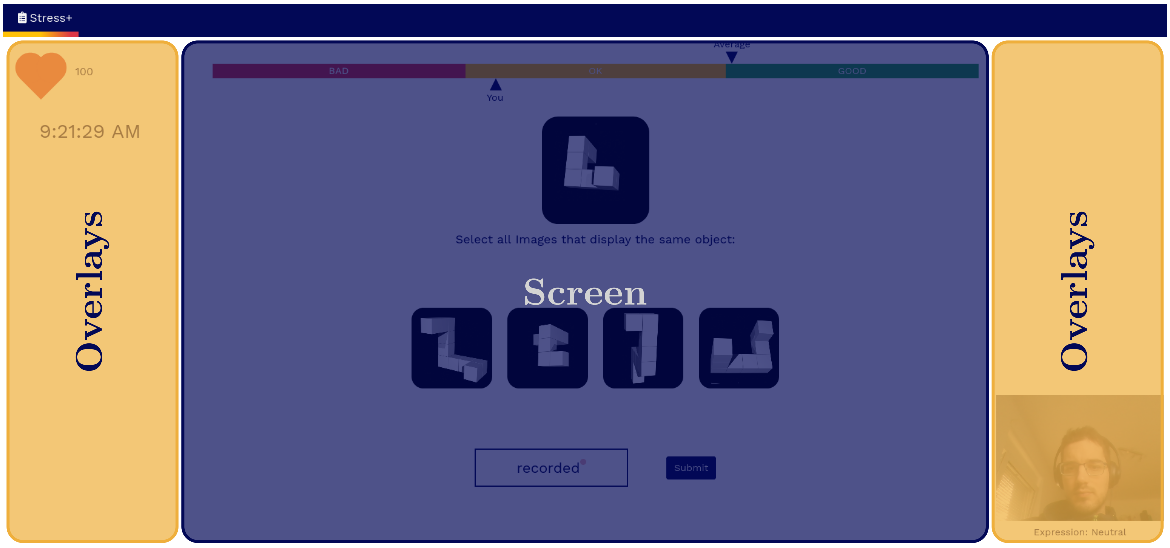


Figure 3.8: Schematic layout of the new user interface

3.1.4 Modifications of the Architecture

The *Executor* is also part of the most significant changes to the code base. To improve the quality of the application's visual appearance, it was essential to update the *Executor* and implementation of generic Screens and Overlays. In the original version, the Screens were functional components and processed the entire display area. This concept had some negative aspects, such as duplication of code. Furthermore, each Screen also offered its own design and the layout of the user interface was not consistent, but significantly varied between Screens. It was, therefore, necessary to re-implement the *Executor* from scratch.

The first step was to consider a new design that would give the modules as much flexibility as possible and could be used for each use case on the screen without making things cluttered. Figure 3.8 shows the schematic layout of the new design with the navigation bar at the top, the optional progress bar below, a main Screen Container in the middle and the two Overlay

Containers on the sides of the screen. Unlike the previous version, this layout restricts the Screens to the central area, keeping the basic design consistent as each sScreen only has to manage this container. Before introducing the Overlay Containers, the Overlays could be placed in each of the four corners and on the four sides. The problem with this solution was that two Overlays overlapped when placed in the same location. This issue was solved by introducing fixed containers in the new implementation. Now the Overlays fill these two containers, starting at the four corners of the monitor along the vertical sides. When two overlays share the same position, they are placed one after the other vertically towards the centre of the container. If a container is too crowded to display all overlays at once, the entire container becomes scrollable. During this redesign, features such as the progress bar and audio playback system were changed from being implemented individually on each screen to being delivered uniformly by the Executor.

This whole unification process also led to introducing a so-called parent component for both types of modules. These new components contain the interfaces for the interaction between a module and the Executor that each module must provide. For screens, these are mainly the methods for passing the progressed time information to the Executor, controlling the audio playback, and handling the frozen state. This state can be set by the Executor to prevent screen content from being re-rendered. Possible use cases for this include simulated screen freezes in the application. Each screen component needs to extend this class, and existing screens have been adapted to this new architecture. Since overlays are simpler than screens, this additional component currently only handles the frozen state. Also, an additional registry for the overlay position was removed with the changes mentioned above. This registry controlled the possible overlay positions on the screen in the previous architecture, but it was no longer necessary with the new and more flexible architecture.

The changes made to the back-end were adding properties to the database schema for storing the participant ID and demographic information.

3.1.5 Screen and Overlay Modules

This work was not just limited to behind-the-scenes improvements but also contained the development of new modules. Due to the fact mentioned above that the first version of the stress+ application was designed to provide the MIST protocol as remote, web-based version, the selection was limited to the following six screens:

- **Start:** Welcome screen for each test protocol
- **End:** End screen for each test protocol
- **Message:** Displays a message and disables the ”Continue” button for a specified time
- **Wait:** Displays a message and automatically continues after a specified time
- **Survey:** Integrates external websites for surveys
- **Math Test:** Sets mental arithmetic tasks similar to MIST

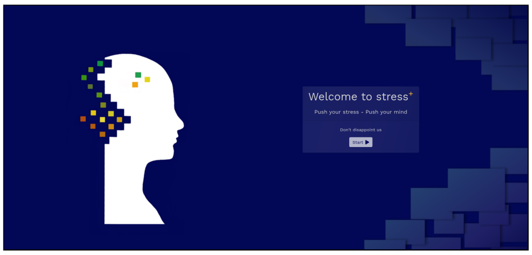
In this selection, only the maths test screen was designed to induce a stress reaction. For that reason, within this thesis alternative stressors were explored and implemented into modules. The first two new modules were usability modules — an *EnterID* screen to assign test runs to participants and a form to collect demographic information. The next step was to think about how to move away from the existing standardised stress protocols and, hopefully, reduce the habituation effect. The chosen solution is to provide the opportunity to try out new combinations of already known stressors. Therefore, the decision was made to select stressors from the list generated during the literature review (Figure 2.1) and implement them into the application.

- **EnterID:** Formular for entering a participant ID
- **Information Form:** Form for demographic information
- **Empty:** Blank pause screen that does not display a message
- **Mental Rotation:** Shows a reference picture next to a random selection of pictures with the same and similar objects from different angles; the participant has to select all pictures showing the same object as in the reference picture

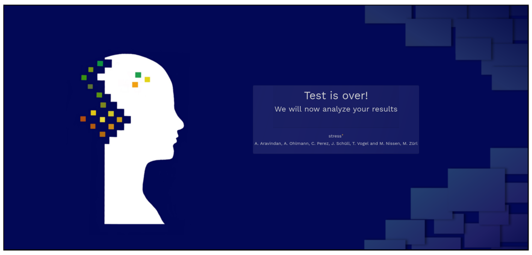
- **Serial Subtraction:** Asks the participant to count down in fixed steps from a starting value; similar to the serial subtraction task in the TSST
- **Video:** Shows an aversive video with the goal of putting the participant in a state of emotional stress
- **Speech:** Asks the participant to give a speech on a specific topic; Can react to pauses during the speech
- **Spatial Stroop:** Displays arrows in a square field; the participant must press the arrow key either according to the position or the direction of the arrow [Wüh07]
- **Numerical Stroop:** Displays two numbers whose size and value do not match; the participant must press the arrow keys according to either the larger representation or, the larger value [Hen82]
- **n-Back:** Displays one letter at a time; the participant must press a button when the letter matches the n-th predecessor [Kir58]

Figures 3.9 and 3.10 keep sample screenshots for each module. Furthermore, by adding these new modules, the types of stressors have also been increased. With these newly added moduls, the stress+ application now contains stressors for each different category as defined in Chapter 2. For example, emotional stress can now be induced by showing aversive videos to the participants. Nevertheless, categories such as cognitive stress have been expanded to include various Stroop tasks, Mental Rotation and n-Back tasks. Each task is configurable with different difficulty levels to adjust the cognitive challenges. In addition, all modules can provide additional psychosocial stress in the form of rating feedback, which is provided through immediate feedback on a given response in the form of ”Correct”, ”Incorrect”, or ”Timeout” if the respondent has not answered within the time limit. However, the Speech module focuses exclusively on social-evaluative stress. In addition to the already stressful situation of delivering a speech, the Speech Module detects when the respondent stops speaking for a predetermined time and prompts them to continue their speech. This reaction improves the sense of social threat.

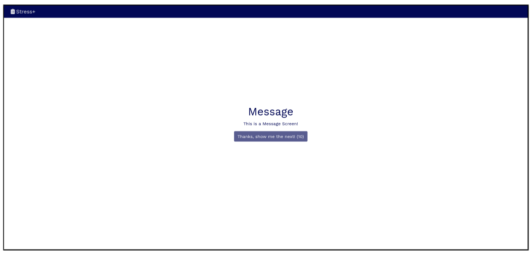
While the number and variety of screen modules increased significantly, this work added only one overlay. This new overlay module allows researchers to show participants a pre-recorded video of a jury watching the subjects on a monitor, creating the illusion of being



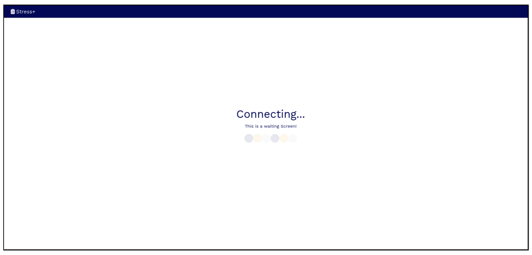
(a) Start



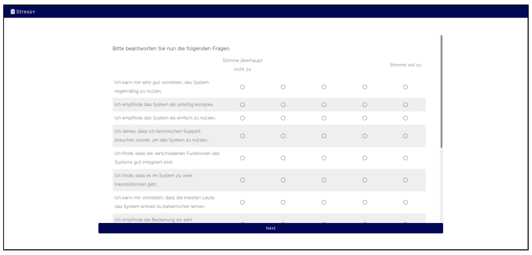
(b) End



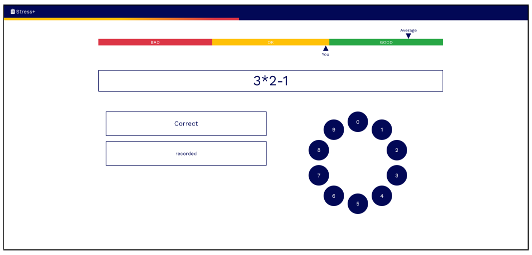
(c) Message



(d) Wait

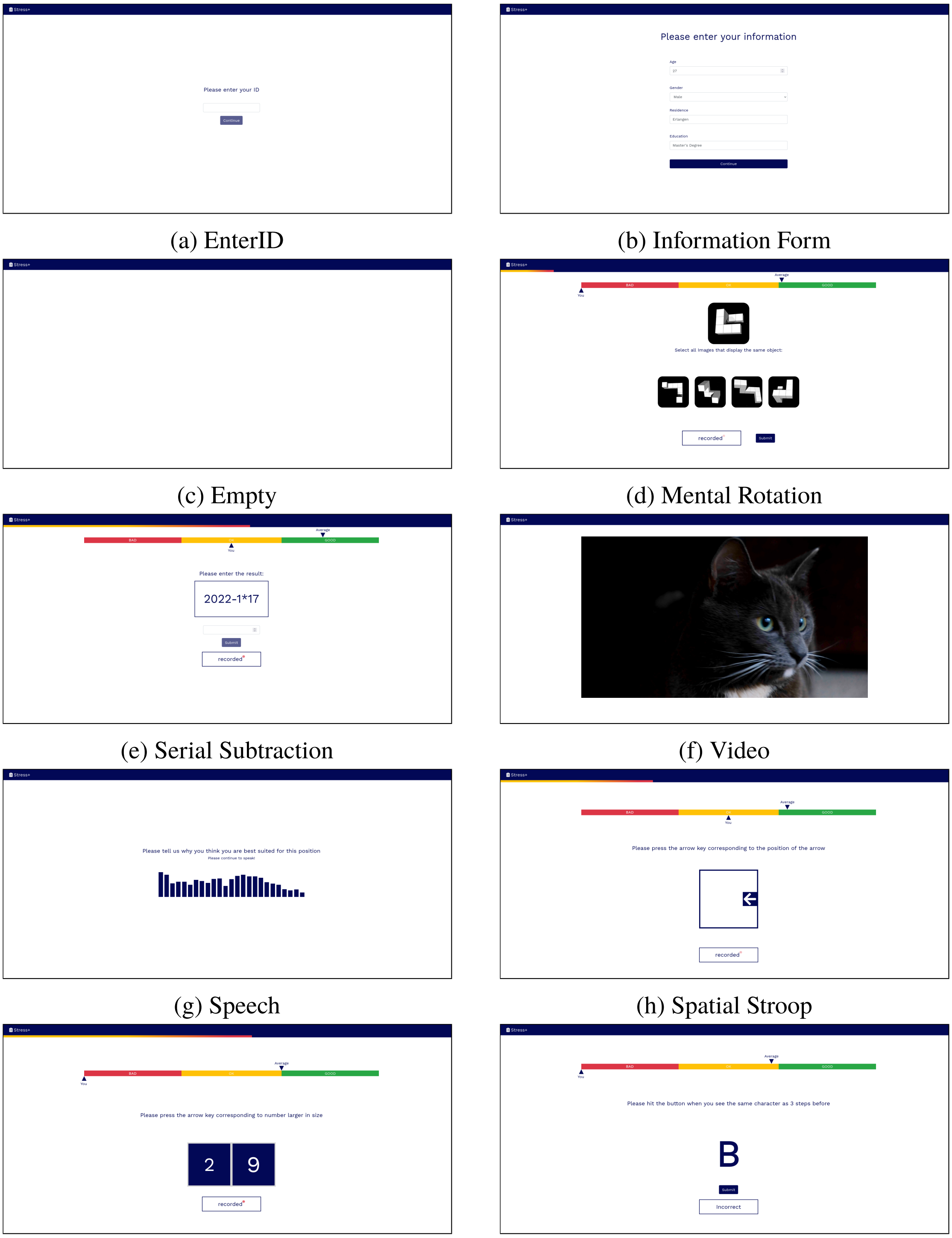


(e) Survey



(f) Math Test

Figure 3.9: Screenshots of already existing Screen Modules



(a) EnterID (b) Information Form
(c) Empty (d) Mental Rotation
(e) Serial Subtraction (f) Video
(g) Speech (h) Spatial Stroop
(i) Numerical Stroop (j) n-Back

Figure 3.10: Screenshots of added Screen Modules

watched while performing the task, thus inducing a social-evaluative threat. This overlay is meant to reinforce the feeling of being watched and thus create a social evaluation threat. In addition, a great deal of effort was spent on improving the existing overlays, mainly done via the modifications in Section 3.1.4. For example, the screen freeze became much more authentic. The chatbot was also improved by adding categories to differentiate between various scenarios. For example, it can give negative feedback, remind participants to ensure a stable internet connection, or inform them that they are currently only in the control group of the experiment. In addition, the available webcam overlay now includes simple but effective recognition of facial emotions. The recognition is done via the JavaScript API face-api.js[4]. The detected emotions are shown to the user below the video stream. The intention was to reinforce the feeling of being analysed and observed. Finally, the application currently contains six overlay modules, which are also shown in Figure 3.11:

- **Current Time:** Displays the current time
- **Screen Freeze:** Leads to a complete freeze of the user interface and a message about the failure of the internet connection
- **Webcam:** Displays the participant's webcam stream; also performs facial expression recognition and displays it to the participant
- **Heart Beat:** Displays a fake heart rate graph to perhaps increase the actual heart rate of the participant
- **Chatbot:** Contacts the user in a pop-up chat window about a given topic
- **Jury Video:** Shows a previously recorded video of the jury in the control room

[4]https://github.com/justadudewhohacks/face-api.js/

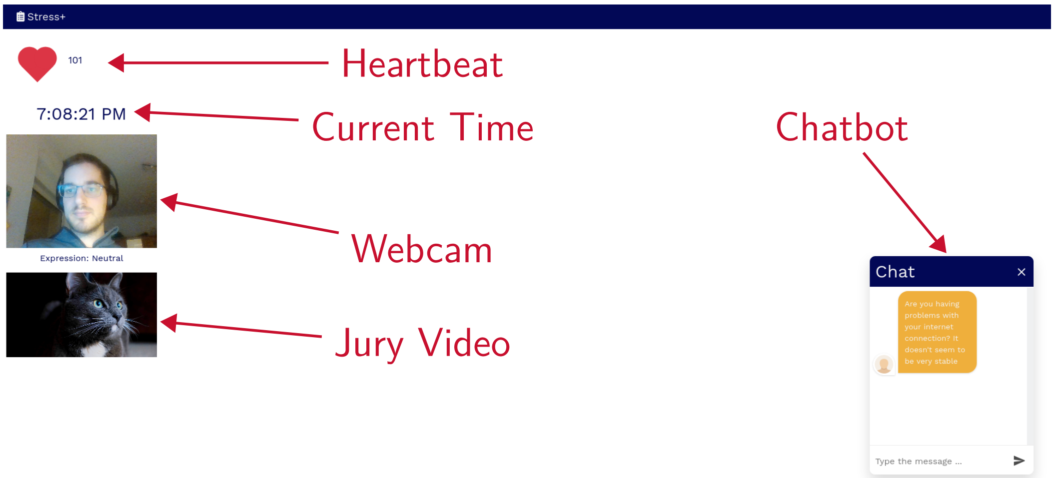


Figure 3.11: Screenshot showing each Overlay Module

3.2 Evaluation

The goal of this evaluation study was to investigate the usability of the application. The main focus was on the researcher interface, which is the more complex part of the user interface. The aim was to assess how easy and effective it is for researchers from different professional backgrounds who have never used the application to interact with the application and create their own stress test pipelines

3.2.1 Study Design

The entire study took place in individual sessions in one room of the MaD Lab. Each participant used the same computer to obtain comparable data. The computer was a Lenovo Thinkpad E570 running Microsoft Windows 10 and the version 97.0 of Mozilla Firefox.

After participants entered the room and took their seats, they were given a short introduction about the purpose of the application and the study's scenario. In the scenario, the participants took on the role of researchers who wanted to conduct stress tests. However, due to contact restrictions during the pandemic, their participants are not allowed into the laboratory for stress tests. For this reason, they decided to use the proposed tool. Before they started, the participants learned that each task's timing and mouse logging would begin as soon as they opened the editor of the web application. Consequently, they were allowed to read the instructions before opening the editor. Before they started, they had the opportunity to ask any final questions, as no further questions were allowed or help offered during the task.

In total, they had five tasks to solve. The instructions for the tasks are attached in the Appendix B. The tasks contained an increasing level of difficulty. In Task 1, participants created a simple test pipeline with an additional Screen and individual settings such as total test time and answer timeout. The first task still contained a step-by-step guide to familiarise the participants with the tool. In Task 2, they then had to add an Overlay and place it in the right corner. In the third Task, they had to create a protocol consisting of a *Spatial Stroop*, an *n-Back Task* and a *Speech*, with breaks in between each. The protocol was also meant to include several *Heartbeat Overlays* and a *Webcam Overlay*. The prescribed configurations were also included in the task instructions but not step by step as in the previous two tasks. Task 4 was a transfer exercise: Participants were given a brief description of the original MIST protocol and how it is usually conducted in-vivo the laboratory. Their task was then to transfer the MIST protocol into a stress+ test pipeline. To be successful, they had to understand the mapping of the different elements in the MIST protocol to their correct counterparts in the tool and configure them correctly. In the last Task 5, they had to import a given protocol pipeline from a configuration file and then run it once to get a brief impression of the executor's user interface.

Capturing actions and checking the completion of tasks was fully integrated into the editor page for this study. So no additional software was used for the study. During the execution of each task, the system recorded the following data:

- Task Completion Status
- Timestamps for Start and End of each Task
- Mouse Movements
- Position of Mouse Clicks
- Number of Saving Attempts
- Stress Test Logs of Failed Saving Attempts

After completing the Tasks, the participants had to complete a final survey. The survey aimed to complement the task performance measurements with information about the participants and self-reported measurements about the workload of the tasks and the assessment of the application's usability. Therefore, in addition to general questions about age, gender, education, impairments of the visual system, and self-reported technical and psychological backgrounds, this survey also included four established questionnaires. The first questionnaire in the survey was the System Usability Scale (SUS) [Jor96]. It assesses the general usability. Individually perceived workload during the tasks was assesed by the Nasa Task Load Index (TLX) [Har86]. To evaluate the overall attractiveness of the application, the AttrakDiff Questionnaire [Has03], and the User Experience Questionnaire (UEQ) [Lau08] were used. In the last part of the survey, the participants were asked in open questions what they particularly liked, what they did not like, or what other feedback they had regarding the application.

Chapter 4

Results

A total of 18 people participated in this study. Participants were members of the lab who were not involved in this project and had no previous experience with the tool. In addition, people with a known background in psychology were recruited by email as they were more in line with the intended users of the software. Divided by gender, there were 15 female and 3 male participants. Their ages ranged from 23 to 31 years, with an average of 26.0 years. 7 of the participants reported eye problems in the form of glasses and contact lenses. Thus, there were no other medical conditions that would lead to exclusion from the study. The highest level of education ranged from high school graduation (3) to bachelor's degree (6), master's degree (8), and completed doctoral degree (1). One participant reported not being a native German speaker. In the comparison between psychologists and engineers, exactly half of the participants rated their knowledge in psychology higher than in engineering. The other half rated it the other way around. Therefore the participants were divided into two equal groups of 9 participants: Psychologists and Engineers.

4.1 Task Execution

Figure 4.1 shows the number of participants who successfully solved the corresponding Task. First of all, it is noticeable that there is no difference between the two groups in terms of the completion rate. Considering that both groups were the same size, they achieved an identical completion rate.

When comparing the individual Tasks, the results show that only two participants did not

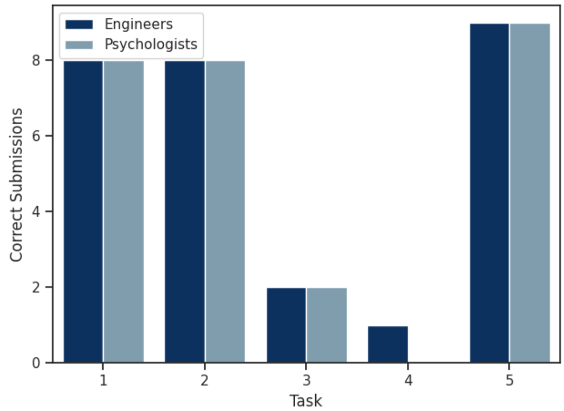


Figure 4.1: Group-wise number of successful submissions per Task

solve the first Task. Subsequent reviews of the submissions showed that both participants had problems changing the test's name. One of them had this problem with all Tasks, so all submissions were classified as incorrect. Nevertheless, the review also revealed that both participants had correctly configured the rest of the protocol as required. In addition, one of them fixed the problem but changed the required configuration during the troubleshooting so that the submission remained incorrect even after the correct naming. In the end, Task 1 has a completion rate of *0.89*. Overall, it took an average of *4:14 Minutes* and *4.39 Attempts* to complete the Task. However, according to Figure 4.2 the differences between the two groups in this Task are relatively small.

For Task 2, the results deliver a similar pattern. Here again, the number of correct entries was identical in both groups. A total of 16 participants successfully solved Task 2. The participants who did not solve this Task were the same as in Task 1. A review of their submissions showed that one participant created a correct Overlay Pipeline but still could not figure out how to save the test protocol in the required naming. The other did not use the copy function as anticipated and created incorrect configurations in the Screen Pipeline. As a result, Task 2 has the same completion rate as Task 1 of *0.89*. The correct submissions took only *1:29 Minutes* and *2.0 Attempts* on average. Psychologists completed this task slightly better than engineers due to fewer attempts. They finished the task with an average of *1.45 Attempts* compared to *2.56 Attempts* in the engineers' group. Consequently, the number of recorded mouse actions was also lower.

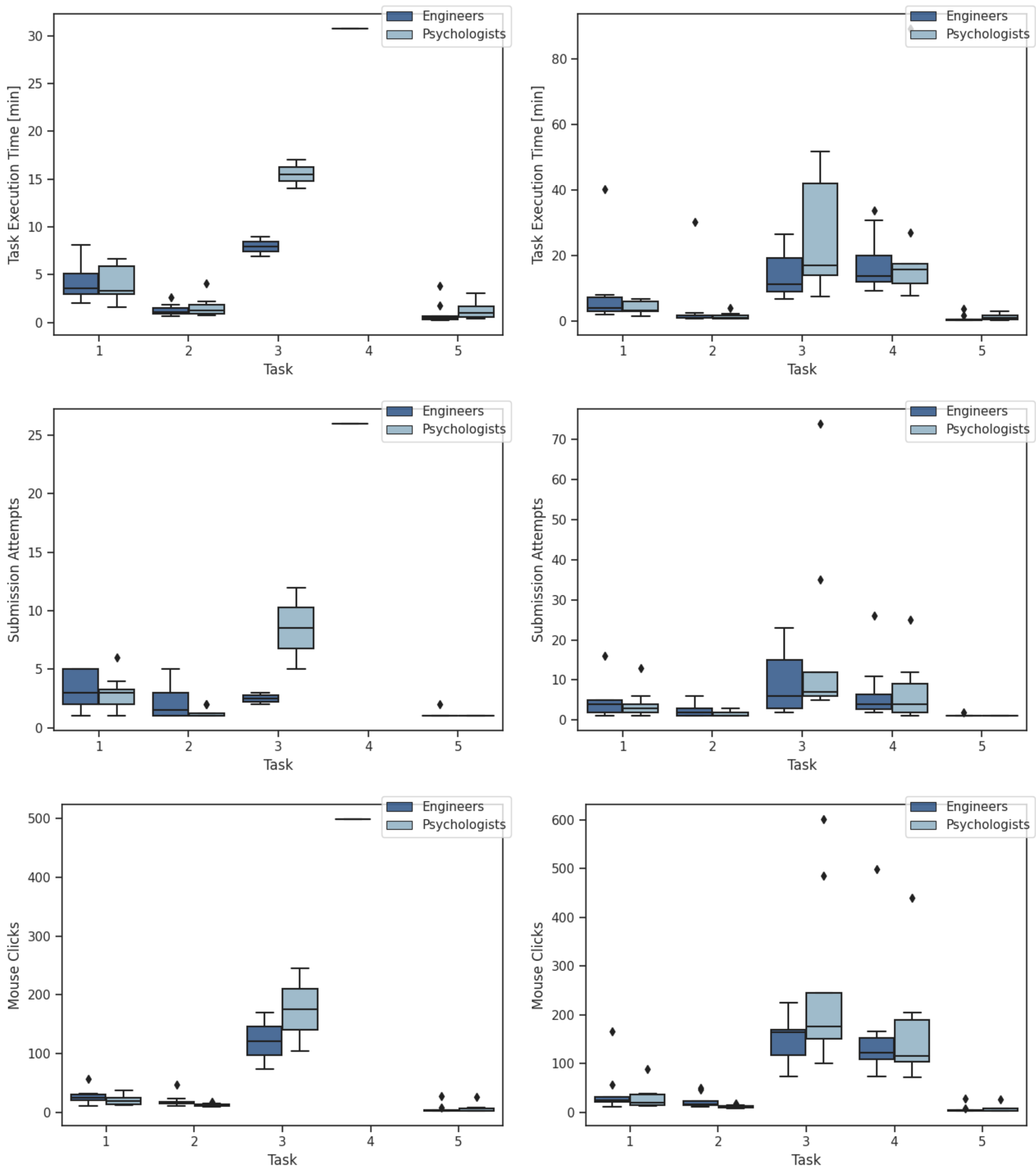


Figure 4.2: Comparison of recorded statistics: correct submissions (left) and total submissions (right)

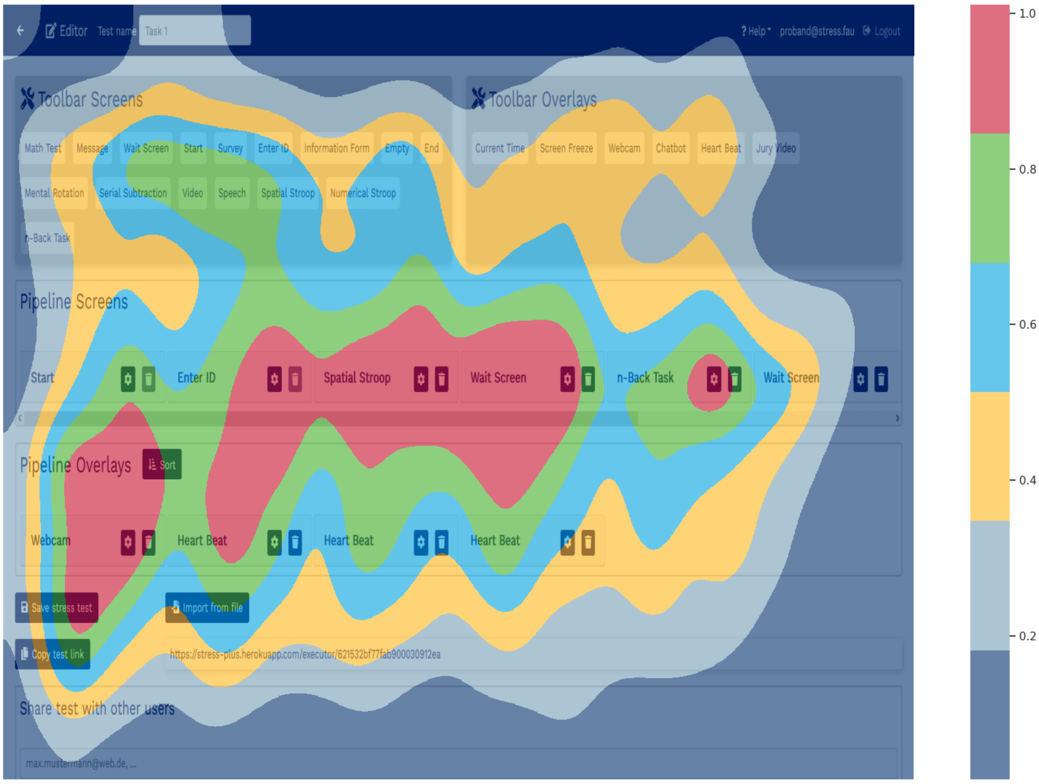


Figure 4.3: Heatmap of mouse movement for Task 3

Task 3 was then the first Task with a higher level of difficulty. This time majority of participants could not solve the Task. Interestingly, the correct solutions are again evenly distributed between the two groups. However, the number of successful submissions drops significantly to just four. And as the complexity of the Task increased, so did the completion time and the number of clicks required. The post-trial review revealed that the most common reason for failure was a minor configuration error of not limiting the difficulty of the Spatial Stroop Task to the arrow direction. This mistake was one of the reasons for failure in 11 of the 14 incorrect final submissions. Other errors included not having a Webcam Overlay, adding unnecessary extra Screens that were not needed, or using Emtpy Screens instead of Waiting Screens. The heatmap of mouse movements in Figure 4.3 confirms these results by showing no movements far off the areas required to solve the Task. However, on the other hand, it also shows that many participants used the Empty Screen, as the number of recorded mouse movements in this area is high. Moreover, it should be noted here that in one final submission,

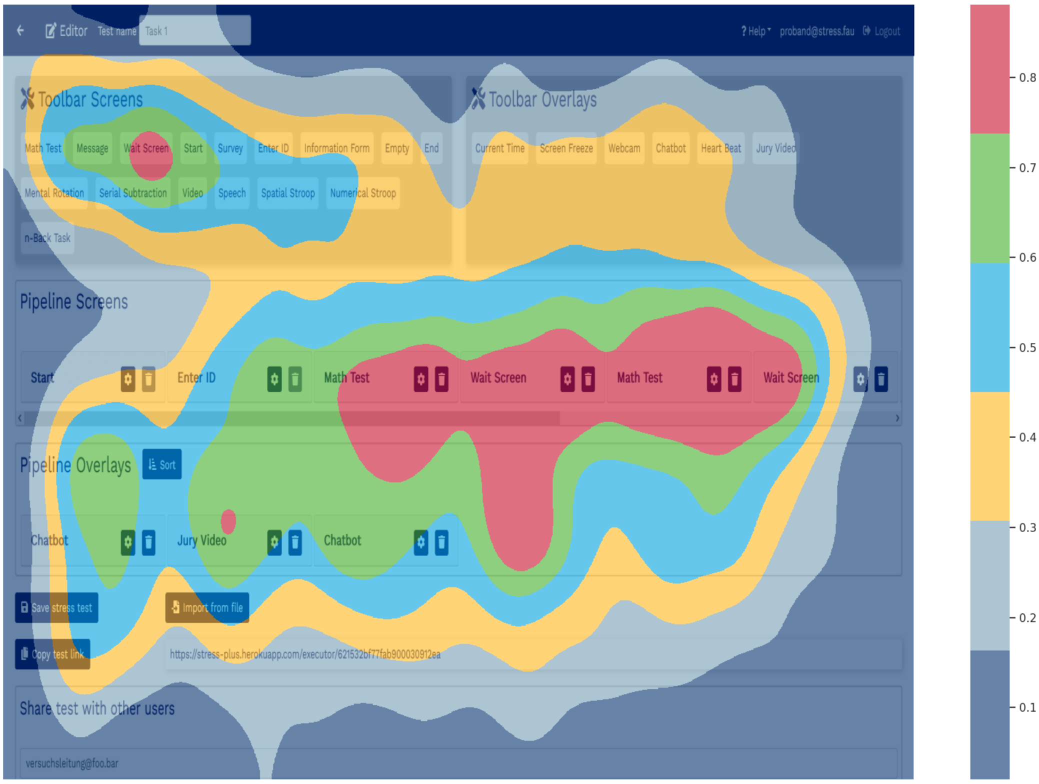


Figure 4.4: Heatmap of mouse movements for Task 4

the only error was using Empty Screens instead of Waiting Screens, while all other parts were correct. Another participant inserted instructional text for the test participants before each of the three Task Screens. It was appropriate and a good idea but failed the check for unnecessary items in the pipeline. Since the Task's instructions did not specify that Waiting Screens should be used instead of Empty Screens, nor did they forbid adding extra Screens, it would have been reasonable to label these two submissions as correct. Nevertheless, to keep the recorded data consistent, this is not done. Thus, on average, the successful solutions were made in a mean of *13.61 Attempts* and *11:44 Minutes*. In contrast to the first two tasks, this task was solved faster and with fewer failed attempts by the two correct solutions from the group of engineers. However, if we compare all participants, this effect is reduced, and the number of attempts and clicks does not differ much between the groups.

For Task 4, only one participant from the engineering group managed to solve it within the study. This participant needed *26 Attempts* and about half an hour (*30:45 Minutes*). The most common mistake made by the others was not adding a chatbot or only adding one for negative feedback and omitting the second one for the control phase. Similar to Task 3, the post-study review showed that two incorrect final submissions were almost correct. One of them contained unnecessary Current Time Overlays, and the other utilised Empty Screens instead of waiting screens, while both configured the rest correctly. However, according to the heatmap of mouse movements (Figure 4.4), this was less of a problem than it had been in Task 3, as the system recorded less activity over and around the Empty Screen item in the Screen Toolbar. Since only one group made a correct submission, we cannot compare performance group-wise. Comparing all the participants, we see that both groups are again close. Also, as we can observe in the heatmap of mouse movements, no one tried to find help in the help menu of the tools. Nevertheless, this was just the application's standard menu and would not have offered any help in solving the Tasks.

Task 5 was indeed the easiest Task and was solved by all participants with a total of only one failed attempt. Therefore, the average completion time of the Task was also low at only *1:03 Minutes*. Given that the correct solution did not require a change in naming, even participants who had previously failed in naming managed to solve this Task.

4.2 Questionnaires

After finishing the Tasks the participants filled out a final survey, containing the four established questionnaires SUS, NASA TLX, AttrakDiff, and UEQ. The following sections contain the individual results for each of these questionnaires.

4.2.1 System Usability Scale

On the SUS the application scored a total mean value of *60.56*. When we compare the group results, it turns out that the group of psychologists rated the overall usability with an average score of *66.11*, compared to *55.00* in the engineers' group. As Figure 4.5 shows, the range of scores is smaller among the psychologists than for engineers. At the upper end of the scale, one engineer and two psychologists rated more than 85.5 points, which corresponds to "Excellent" usability. [Ban09] But at the lower end of the scale, two engineers scored significantly lower than any psychologists.

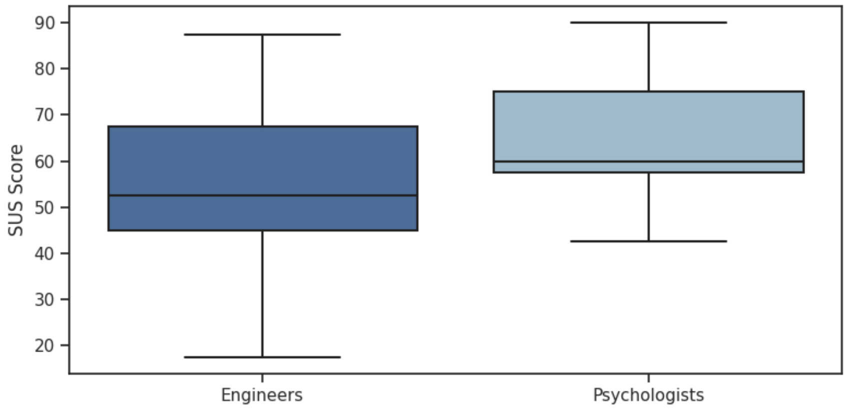


Figure 4.5: Results of System Usability Scale

4.2.2 NASA TLX

The analysis of the NASA TLX questionnaire resulted in a mean overall task load rating of *42.1*. It is remarkable that although the psychologists rated the usability of the SUS higher, they also rated their perceived task load higher (Figure 4.6). On average, they rated their perceived task load *18%* higher than the engineers. However, comparing the individual subscales of the questionnaire (Figure 4.7), the engineers consistently rated their perceived task load lower, except for physical demands. The smallest differences between the groups are in categories of Physical Demands and individual performance rating. In the other categories, the absolute value of the difference is almost stable.

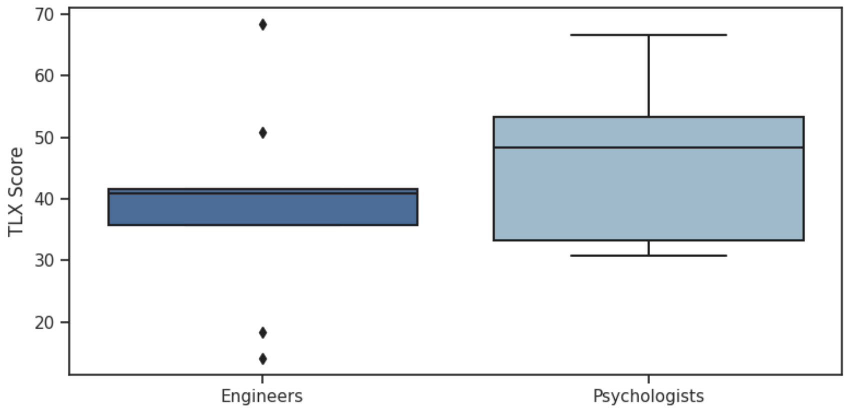


Figure 4.6: Results of NASA TLX

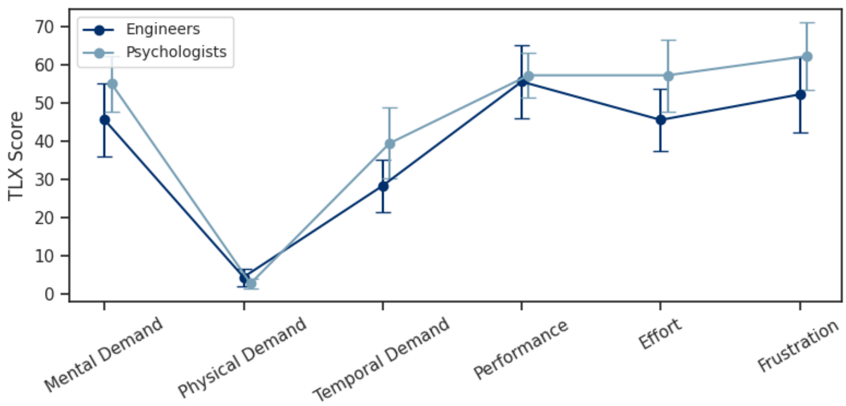


Figure 4.7: NASA TLX Subscale Scores

4.2.3 AttrakDiff

Group	PQ	HQ-I	HQ-S	ATT
Engineers	0.98	1.56	1.65	1.38
Psychologists	1.37	1.70	1.76	1.73
Total	1.17	1.63	1.71	1.56

Table 4.1: Results of AttrakDiff

Since AttrakDiff does not give a total score, Table 4.1 shows the results for the four subscales. This questionnaire analyses the *Pragmatic Quality (PQ)*, the *Hedonic Quality*, divided into *Identity (HQ-I)* and *Stimulation (HQ-S)*, and the *Attractiveness (ATT)* of the application. The results show that here, once again, the psychologists rated the individual subscales higher than the engineers.

4.2.4 User Experience Questionnaire

Group	ATT	PQ	HQ
Engineers	0.76	0.46	1.04
Psychologists	1.33	1.06	1.44
Total	1.05	0.76	1.24

Table 4.2: Results of UEQ

The UEQ was the final questionnaire of the survey. Similar to AttrakDiff, it assesses *Pragmatic Quality*, *Hedonic Quality* and *Attractiveness*. Table 4.2 shows the final results. However, we first need to determine the six subscales shown in Figure 4.8 to calculate these results. The average values of *Dependability*, *Efficiency*, and *Perspicuity* form the *Pragmatic Quality*. The mean score of *Novelty* and *Stimulation*, on the other hand, reflect *Hedonic Quality*.

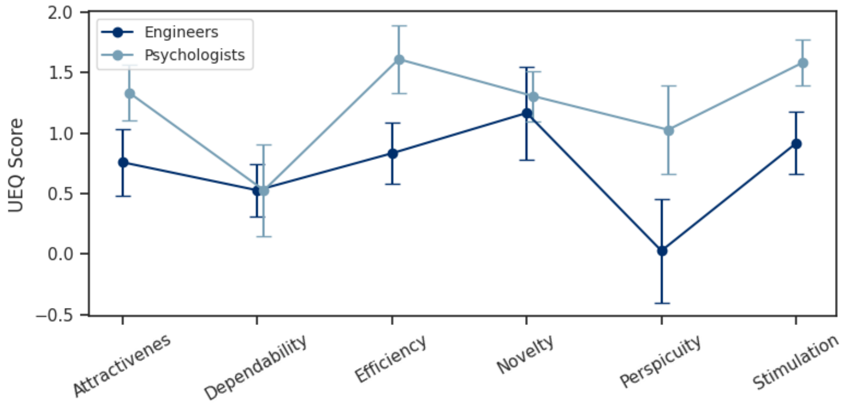


Figure 4.8: UEQ Subscale Scores

Chapter 5

Discussion

Before discussing the study results in terms of the research goals of this thesis, it is essential to mention a few things that could be considered relevant regarding the study design. As these problems could be considered as limiting factors or influences on the study. The first issue is the clarity of the Tasks. Immediately after the study as well as in the open text responses in the final survey, many subjects stated that the Tasks were not always understandable, often confusing, and therefore unnecessarily complex. This issue applies in particular to Task 3 and Task 4. Participants often did not know what was expected in the Tasks and were confused about which Screen Module to use when several options made sense. More specifically, some described their troubles with Empty and Waiting Screens, which was already noted in the results section. Some others added that the confusing task instructions also led to frustration and negatively influenced their opinion of the application. Since this had already been raised during the study, it would have been possible to adjust the task instructions. However, this was deliberately not done as it would have required the exclusion of all participants who had already completed the study using the original instructions. The second criticism that participants voiced was that they only received a standard error message when they saved the wrong pipelines for a task. And that this in turn led to further frustration. However, this happened on purpose to prevent participants from just following the error message without going deeper into the application and doing the expected transfer thinking themselves.

So much for that, but if we now look at the results of the study with regard to the goals to be evaluated, we can see both positive and negative aspects. As a reminder, this study aimed to assess the usability and the efficiency of the application, as well as the differences between the two groups of participants when interacting with the application. The collected results show that the user experience still needs to be improved. This deficiency is particularly evident in the high task completion times and the number of failed attempts. Especially in Task 3 and Task 4, participants spent a lot of time without necessarily solving the task. This can therefore not be considered efficient and underlines the need for improved user guidance. Moreover, the number of attempts in some cases looks more like trial and error than exact knowledge of what to do and how to do it. In Task 3, for example, one participant made over 70 attempts and still was unable to solve the task. Of course, this may be partly due to the above-mentioned lack of clarity in the task instructions, but not entirely. In addition, the participants' answers in the open questions of the survey included requests and suggestions for improvements to the user interface, which further underline the need. The most frequent criticism was the lack of a more detailed description of the individual modules so that users did not know what each module did. The majority of participants would have welcomed additional information such as texts or even some kind of preview of the individual modules within the application. Other suggestions that would improve usability were a copy function for Modules in the Pipeline and sorting of items within the toolbox. Indeed, the Modules in the Toolbar were not sorted by function or alphabetically. The order in this version was more in the sequence of the development of the Modules. The survey results also underline these points raised so far. When looking at the scores in the questionnaire, the results do not show optimal usability either. For example, the application only scored mean score of 60.56 in the SUS. Other studies and benchmarks classify the usability of an application as "Good" if it achieves a score of 71.4 or higher [Ban09]. So with this score, the usability of the application is only in the middle range of "OK". Furthermore, the overall user experience, which is closely related to the usability of a system, also has some deficits. Looking at the NASA TLX questionnaire, the high perceived workload is also striking. Although there are no references or benchmarks for comparison, the scores given by some participants for effort and frustration are striking. For instance, two participants selected the highest possible score for Frustration in the NASA TLX questionnaire. Such high scores are not acceptable when aiming for a good user experience and usability. Furthermore, and probably as a result, the user experience was judged to be unclear, so that the application was also unable to generate a high level of user trust. The mean scores< of UEQ subscales for Perspicuity and Dependability are clearly in the "below average" category compared to the benchmarks for web applications given in the UEQ Handbook [Sch15].

Nevertheless, there are also positive findings from the results of this study. Even though usability did not receive the best ratings, many of the application's features were positively highlighted in the question about what users like about the application. For example, the ability to move items via drag & drop and the elimination of unnecessary elements on the user interface have been well received. Regarding the second research question about the difference in interaction with the application between engineers and psychologists, it turned out that there are hardly any differences in the use of the application and the completion of tasks. For example, looking at the completion rate, both groups had exactly the same number of correct submissions in four out of five tasks, while being equal in size. Only task 4 differed here, with this generally being solved by only one person. Slight differences become visible when comparing the recorded data for Task 3. The psychologists spent more time on this task than the engineers, regardless of whether they were able to solve the task or not. They also consequently did more attempts and more mouse clicks. However, when comparing the other tasks, this difference is much smaller. So this difference is not significant enough to draw any conclusions.

Nevertheless, the evaluation of the questionnaires' results shows a more significant difference between the groups. It should be noted here that the group of psychologists on average gave better ratings on most scales compared to the engineers' group. For example, the psychologists' average score for perceived usability on the SUS was well above the overall average score. Other scales with remarkable differences between the two groups were the Perspicuity subscale in the NASA TLX as well as the Attractiveness and Pragmatic Quality subscales in AttrakDiff and UEQ. This effect is probably due to the different personal requirements and previous experience with software tools in the two groups. It is quite likely that the engineers tend to evaluate applications more critically than psychologists due to their technical experience and personal standards. In addition, the psychologists' experience with the general topic of stress testing could be a reason for the large difference in the ratings of Perspicuity and the judgement of Efficiency. This is because, due to their experience, many of them knew how much effort laboratory stress tests require and might therefore have rated efficiency higher. Engineers do not normally have this comparison. On top of that, the categories stimulation and novelty are particularly noteworthy. According to the UEQ Handbook, the ratings, especially by the psychologists, for these categories are classified as ”Good” and ”Excellent”, supporting our overall goal of creating a novel tool that people enjoy using for work. This supports the overall goal of this thesis of creating a novel tool that people enjoy using for their work. Considering that psychologists are the intended users, the results are encouraging and show that this tool can be widely accepted. Moreover, this study has shown that participants are open to it, even if it has not yet provided the best user experience.

Chapter 6

Conclusion and Future Work

In summary, this thesis has significantly advanced the stress+ platform. In the process of this improvement, ten new Screen Modules have been developed and added, most of which feature new stressors from different categories. So the platform now contains a total of 16 Screen Modules and 6 Overlay Modules. Furthermore, major changes have been made to the architecture. As part of these changes, the entire Executor was revised and a new concept for positioning of Overlays and Screens was developed. The subsequent study examined the usability and effectiveness of the user interface, as well as whether there was a difference between people with technical experience and those with psychological experience in terms of interaction with the application. In this study, participants had to complete five different tasks and rate their impressions in a final survey. The analysis of the results of the study revealed some major deficits in the usability of the application. At the same time, the results also provided feedback for improvements and indicated that there is a demand for this novel and innovative application. These findings were measurable in the average overall ratings of the corresponding scales. However, this judgement is even more pronounced when looking only at the ratings of psychologists, who are the actual target group of this application.

The next steps for the further development of the platform are now to optimise the user interface of the Editor in order to improve the user guidance and user friendliness. These improvements includes in particular the implementation of the functions mentioned in the discussion, such as additional description texts and preview functions for each module. Whether the preview will be a simple screenshot or a real-time preview showing the current combination must be judged on feasibility and practicality.

Beyond that, the platform also needs to be assessed in further research studies. The most important aspect to be investigated is whether the application can reliably induce stress. This is after all the essential function and the main purpose of the application. This will involve careful planning and creation of stress pipelines and subsequent comparison with laboratory test protocols such as the TSST. Therefore, this is certainly the most labour-intensive part of the future tasks, but also the most important. In summary, despite the progress that has been made with the platform and the nevertheless positive results of this thesis, there is still much to be done.

List of Figures

2.1 Stressors used in common stress protocols [Dic04] 4
2.2 User Interface of the MIST [Ded05] 6

3.1 User interface for researchers 11
3.2 User interface for Participants showing a *Mental Rotation Task* and a *Webcam*, a *Heartbeat*, and a *Current Time Overlay* 12
3.3 Application Architecture 13
3.4 Routes of the REST API 15
3.5 Schema of Stress Test Collection 16
3.6 Schema of Stats Collection 16
3.7 Schema of Rule Policy Collection 17
3.8 Schematic layout of the new user interface 17
3.9 Screenshots of already existing Screen Modules 21
3.10 Screenshots of added Screen Modules 22
3.11 Screenshot showing each Overlay Module 24

4.1 Group-wise number of successful submissions per Task 28
4.2 Comparison of recorded statistics: correct submissions (left) and total submissions (right) 29
4.3 Heatmap of mouse movement for Task 3 30
4.4 Heatmap of mouse movements for Task 4 31
4.5 Results of System Usability Scale 33
4.6 Results of NASA TLX 34
4.7 NASA TLX Subscale Scores 34
4.8 UEQ Subscale Scores 36

List of Tables

2.1 The Trier Social Stress Test [Kir93] 5

4.1 Results of AttrakDiff 35
4.2 Results of UEQ 35

Bibliography

[Ban09] Aaron Bangor, Philip Kortum, and James Miller. “Determining what individual SUS scores mean: adding an adjective rating scale”. In: *J. Usability Studies* 4.3 (May 2009), pp. 114–123. ISSN: 1931-3357.

[Coh07] Sheldon Cohen, Denise Janicki-Deverts, and Gregory E. Miller. “Psychological Stress and Disease”. In: *JAMA* 298.14 (Oct. 2007), pp. 1685–1687. ISSN: 0098-7484. DOI: 10.1001/jama.298.14.1685. URL: https://doi.org/10.1001/jama.298.14.1685.

[De 21] Joana De Calheiros Velozo, Thomas Vaessen, Jens Pruessner, Ilse Van Diest, Stephan Claes, and Inez Myin-Germeys. “The repeated Montreal Imaging Stress Test (rMIST): Testing habituation, sensitization, and anticipation effects to repeated stress induction”. en. In: *Psychoneuroendocrinology* 128 (June 2021), p. 105217. ISSN: 0306-4530. DOI: 10.1016/j.psyneuen.2021.105217. URL: https://www.sciencedirect.com/science/article/pii/S0306453021000913.

[Dec18] Lambert Deckers. *Motivation: Biological, Psychological, and Environmental*. 5th ed. New York: Routledge, Feb. 2018. ISBN: 978-1-315-17861-5. DOI: 10.4324/9781315178615.

[Ded05] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J. Lupien, and Jens C. Pruessner. “The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain”. In: *Journal of Psychiatry and Neuroscience* 30.5 (Sept. 2005), pp. 319–25. ISSN: 1180-4882. URL: http://www.ncbi.nlm.nih.gov/pubmed/16151536.

[Dic04] Sally S. Dickerson and Margaret E. Kemeny. "Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research". In: *Psychological Bulletin* 130.3 (2004). ISBN: 0033-2909 (Print)\n0033-2909 (Linking), pp. 355–391. ISSN: 00332909. DOI: 10.1037/0033-2909.130.3.355.

[Eag21] David E. Eagle, Joshua A. Rash, Logan Tice, and Rae Jean Proeschold-Bell. "Evaluation of a remote, internet-delivered version of the Trier Social Stress Test". en. In: *International Journal of Psychophysiology* 165 (July 2021), pp. 137–144. ISSN: 01678760. DOI: 10.1016/j.ijpsycho.2021.03.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167876021001318.

[Har86] Sandra G Hart. "NASA Task Load Index (TLX)". In: (1986). URL: https://humansystems.arc.nasa.gov/groups/tlx/.

[Has03] Marc Hassenzahl, Michael Burmester, and Franz Koller. "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität". de. In: *Mensch & Computer 2003: Interaktion in Bewegung*. Ed. by Gerd Szwillus and Jürgen Ziegler. Berichte des German Chapter of the ACM. Wiesbaden: Vieweg+Teubner Verlag, 2003, pp. 187–196. ISBN: 978-3-322-80058-9. DOI: 10.1007/978-3-322-80058-9_19. URL: https://doi.org/10.1007/978-3-322-80058-9_19.

[Hen82] Avishai Henik and Joseph Tzelgov. "Is three greater than five: The relation between physical and semantic size in comparison tasks". en. In: *Mem Cogn* 10.4 (July 1982), pp. 389–395. ISSN: 1532-5946. DOI: 10.3758/BF03202431. URL: https://doi.org/10.3758/BF03202431.

[Jön10] Peter Jönsson, Mattias Wallergård, Kai Österberg, Åse Marie Hansen, Gerd Johansson, and Björn Karlson. "Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the Trier Social Stress Test: A pilot study". In: *Psychoneuroendocrinology* 35.9 (2010), pp. 1397–1403. ISSN: 03064530. DOI: 10.1016/j.psyneuen.2010.04.003.

[Jor96] Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. *Usability Evaluation In Industry*. en. Google-Books-ID: IfUsRmzAqvEC. CRC Press, June 1996. ISBN: 978-0-7484-0460-5.

[Kal02] Madhu Kalia. "Assessing the economic impact of stress[mdash]The modern day hidden epidemic". English. In: *Metabolism - Clinical and Experimental* 51.6 (June 2002). Publisher: Elsevier, pp. 49–53. ISSN: 0026-0495, 1532-8600. DOI: 10.1053/meta.2002.33193. URL: https://www.metabolismjournal.com/article/S0026-0495(02)43792-4/abstract.

[Kän06] Roland von Känel, Brigitte M. Kudielka, Daniel Preckel, Dirk Hanebuth, and Joachim E. Fischer. "Delayed response and lack of habituation in plasma interleukin-6 to acute mental stress in men". en. In: *Brain, Behavior, and Immunity* 20.1 (Jan. 2006), pp. 40–48. ISSN: 0889-1591. DOI: 10.1016/j.bbi.2005.03.013. URL: https://www.sciencedirect.com/science/article/pii/S0889159105000644.

[Kel07] Owen Kelly, Kimberly Matheson, Alejandra Martinez, Zul Merali, and Hymie Anisman. "Psychosocial Stress Evoked by a Virtual Audience: Relation to Neuroendocrine Activity". In: *CyberPsychology & Behavior* 10.5 (Oct. 2007). Publisher: Mary Ann Liebert, Inc., publishers, pp. 655–662. ISSN: 1094-9313. DOI: 10.1089/cpb.2007.9973. URL: https://www.liebertpub.com/doi/10.1089/cpb.2007.9973.

[Kir58] Wayne K. Kirchner. "Age differences in short-term retention of rapidly changing information". In: *Journal of Experimental Psychology* 55.4 (1958). Place: US Publisher: American Psychological Association, pp. 352–358. ISSN: 0022-1015. DOI: 10.1037/h0043688.

[Kir93] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting". In: *Neuropsychobiology*. Vol. 28. Issue: 1-2 ISSN: 0302282X. 1993, pp. 76–81. ISBN: 978-0-87421-656-1. DOI: 10.1159/000119004. URL: 10.1159/000119004.

[Kir94] Clemens Kirschbaum and Dirk H. Hellhammer. "Salivary cortisol in psychoneuroendocrine research: Recent developments and applications". en. In: *Psychoneuroendocrinology* 19.4 (Jan. 1994), pp. 313–333. ISSN: 0306-4530. DOI: 10.1016/0306-4530(94)90013-2. URL: https://www.sciencedirect.com/science/article/pii/0306453094900132.

[Lau08] Bettina Laugwitz, Theo Held, and Martin Schrepp. "Construction and Evaluation of a User Experience Questionnaire". In: vol. 5298. Nov. 2008, pp. 63–76. ISBN: 978-3-540-89349-3. DOI: 10.1007/978-3-540-89350-9_6.

[McE98] Bruce S. McEwen. "Protective and Damaging Effects of Stress Mediators". In: *New England Journal of Medicine* 338.3 (1998). Ed. by Jeffrey S. Flier and Lisa H. Underhill. Publisher: Massachusetts Medical Society, pp. 171–179. ISSN: 00284793. DOI: 10.1056/NEJM199801153380307. URL: http://www.nejm.org/doi/10.1056/NEJM199801153380307.

[Noa19] Hannes Noack, Leandra Nolte, Vanessa Nieratschker, Ute Habel, and Birgit Derntl. "Imaging stress: an overview of stress induction methods in the MR scanner". en. In: *J Neural Transm* 126.9 (Sept. 2019), pp. 1187–1202. ISSN: 1435-1463. DOI: 10.1007/s00702-018-01965-y. URL: https://doi.org/10.1007/s00702-018-01965-y.

[Pet12] Katja Petrowski, Gloria-Beatrice Wintermann, and Martin Siepmann. "Cortisol Response to Repeated Psychosocial Stress". en. In: *Appl Psychophysiol Biofeedback* 37.2 (June 2012), pp. 103–107. ISSN: 1573-3270. DOI: 10.1007/s10484-012-9183-4. URL: https://doi.org/10.1007/s10484-012-9183-4.

[Roh19] Nicolas Rohleder. "Stress and inflammation – The need to address the gap in the transition between acute and chronic stress effects". In: *Psychoneuroendocrinology* (2019). Publisher: Elsevier Ltd. ISSN: 18733360. DOI: 10.1016/j.psyneuen.2019.02.021. URL: 10.1016/j.psyneuen.2019.02.021.

[Rui10] Ana Santos Ruiz, Maria Isabel Peralta-Ramirez, Maria Carmen Garcia-Rios, Miguel A. Muñoz, Nuria Navarrete-Navarrete, and Antonia Blazquez-Ortiz. "Adaptation of the trier social stress test to virtual reality: Psycho-phsyiological and neuroendocrine modulation". In: *Journal of Cyber Therapy and Rehabilitation* 3.4 (2010), pp. 405–415. ISSN: 17849934.

[Sap00] Robert M. Sapolsky, L. Michael Romero, and Allan U. Munck. "How Do Glucocorticoids Influence Stress Responses? Integrating Permissive, Suppressive, Stimulatory, and Preparative Actions*". In: *Endocrine Reviews* 21.1 (Feb. 2000), pp. 55–89. ISSN: 0163-769X. DOI: 10.1210/edrv.21.1.0389. URL: https://doi.org/10.1210/edrv.21.1.0389.

[Sch03] Nicole C. Schommer, Dirk H. Hellhammer, and Clemens Kirschbaum. “Dissociation Between Reactivity of the Hypothalamus-Pituitary-Adrenal Axis and the Sympathetic-Adrenal-Medullary System to Repeated Psychosocial Stress”. en-US. In: *Psychosomatic Medicine* 65.3 (June 2003), pp. 450–460. issn: 0033-3174. doi: 10.1097/01.PSY.0000035721.12441.17. url: https://journals.lww.com/psychosomaticmedicine/Abstract/2003/05000/Dissociation_Between_Reactivity_of_the.18.aspx.

[Sch15] Martin Schrepp. *User Experience Questionnaire Handbook*. Tech. rep. Sept. 2015. doi: 10.13140/RG.2.1.2815.0245.

[Shi16] Youssef Shiban, Julia Diemer, Simone Brandl, Rebecca Zack, Andreas Mühlberger, and Stefan Wüst. “Trier Social Stress Test in vivo and in virtual reality: Dissociation of response domains”. In: *International Journal of Psychophysiology* 110 (2016). Publisher: Elsevier B.V., pp. 47–55. issn: 18727697. doi: 10.1016/j.ijpsycho.2016.10.008. url: http://dx.doi.org/10.1016/j.ijpsycho.2016.10.008.

[Wüh07] Peter Wühr. “A Stroop Effect for Spatial Orientation”. In: *The Journal of General Psychology* 134.3 (July 2007), pp. 285–294. issn: 0022-1309. doi: 10.3200/GENP.134.3.285-294. url: https://doi.org/10.3200/GENP.134.3.285-294.

[Wüs05] Stefan Wüst, Ilona S. Federenko, Elisabeth F. C. van Rossum, Jan W. Koper, and Dirk H. Hellhammer. “Habituation of cortisol responses to repeated psychosocial stress—further characterization and impact of genetic factors”. en. In: *Psychoneuroendocrinology* 30.2 (Feb. 2005), pp. 199–211. issn: 0306-4530. doi: 10.1016/j.psyneuen.2004.07.002. url: https://www.sciencedirect.com/science/article/pii/S0306453004001209.

[Zim19] Patrick Zimmer, Benjamin Buttlar, Georg Halbeisen, Eva Walther, and Gregor Domes. “Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses”. In: *Psychoneuroendocrinology* 101.August 2018 (2019). Publisher: Elsevier, pp. 186–192. issn: 18733360. doi: 10.1016/j.psyneuen.2018.11.010. url: https://doi.org/10.1016/j.psyneuen.2018.11.010.

Appendix A

Acronyms

ACTH adrenocorticotropic hormone

ANS Autonomic Nervous System

ATT Attractiveness

CRH corticotropin-releasing hormone

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

GH growth hormone

HPA Hypothalamic-Pituitary-Adrenal

HQ Hedonic Quality

HQ-I Hedonic Quality - Identity

HQ-S Hedonic Quality - Stimulation

MaD Machine Learning and Data Analytics

MIST Montreal Imaging Stress Task

PQ Pragmatic Quality

SUS System Usability Scale

TLX Task Load Index

TSST Trier Social Stress Test

UEQ User Experience Questionnaire

VR Virtual Reality

Appendix B

Task Instructions

Benutzer: proband@stress.fau **Passwort:** Stress123!

Hinweis: Einstellungen, welche nicht explizit in der Aufgabenstellung vorgegeben sind, können frei konfiguriert werden!

Aufgabe 1:
Öffnen Sie den Editor indem Sie auf der Übersichtsseite auf "+ **New**" klicken.
Erstellen Sie nun eine einfache Stresstest-Pipeline. Diese soll neben den bereits vorhandenen "Start", "Enter-ID" und "End" Screens auch einen sogenannten "Mental Rotation" Task enthalten. Konfigurieren Sie den Test nun nach folgenden Vorgaben (klicken Sie hierfür jeweils auf ⚙ um das Einstellungsmenü eines Screens zu öffnen):

- Auf dem "Enter ID" Screen sollen die Probanden mittels des Textes "Bitte geben Sie Ihren Nachnamen als ID ein" zur Eingabe der Probanden-ID aufgefordert werden.
- Der "Mental Rotation" Task soll 120 Sekunden dauern und die Probanden sollen pro Antwort 3 Sekunden Zeit haben.
- Zwischen den einzelnen Aufgabe sollen 2 Sekunden liegen.
- Der Schwierigkeitsgrad des "Mental Rotation" Task soll einfach gehalten werden. Wählen Sie dafür als Schwierigkeit "Easy".

Speichern Sie den erstellten Stresstest unter dem Namen "Task 1" ab.

Aufgabe 2:
Fügen Sie nun ein Overlay hinzu. Dazu kopieren Sie den in Aufgabe 1 erstellten Stresstest indem Sie auf der Übersichtsseite auf klicken. Öffnen Sie nun die neu erstellte Kopie "Task 1 (Copy)" indem sie auf klicken und fügen Sie das "Current Time" Overlay hinzu.

Konfigurieren Sie das Overlay so, dass es in der oberen linken Ecke angezeigt wird. Das Overlay soll außerdem nur auf dem "Mental Rotation" Screen angezeigt werden. Wählen Sie hierfür sowohl als "Start Screen" als auch "End Screen" den "Mental Rotation" Screen.
Speichern Sie diesen Stresstest nun unter "Task 2" ab.

Aufgabe 3:
Erstellen Sie nun einen Stresstest, welcher aus drei unterschiedlichen Aufgaben mit dazwischenliegenden Pausen bestehen soll. Die Probanden sollen während des gesamten Tests gefilmt werden. Zwischen den Aufgaben soll jeweils eine Pause von 3 Minuten sein.

In der ersten Aufgabe sollen die Probanden einen sogenannten "Spatial Stroop" Test bearbeiten. Dabei werden Pfeile präsentiert und die Probanden sollen, je nach Pfeilrichtung, die entsprechenden Pfeiltasten drücken. Als Zweites soll ein "N-Back Task" mit N=2 durchgeführt werden. Abschließend sollen die Probanden noch 5 Minuten etwas über ihre Freizeit erzählen.

Zudem soll, allerdings nur während den Aufgaben, die Herzfrequenz als Biofeedback eingeblendet werden. Diese soll erst linear von 70 Schläge/Minute auf 130 Schläge/Minute ansteigen. Nach einer kurzen Erholungsphase während der Pause beginnt sie bei der folgenden Aufgabe bei 100 Schläge/Minute und steigt auf 140 Schläge/Minute an. Bei der letzten Aufgabe hat sie sich bei konstanten 100 Schläge/Minute eingependelt.
Speichern Sie diesen Test als "Task 3" ab.

Aufgabe 4:
Im Folgenden wird Ihnen ein etabliertes Stressprotokoll aus der psychologischen Forschung vorgestellt. Ihre Aufgabe wird es anschließend sein, dieses Protokoll so genau wie möglich in der stress+ App zu realisieren.

Die Montreal Imaging Stress Task (MIST) ist ein Protokoll zur Induktion von akutem psychosozialem Stress. Es wurde ursprünglich für den Einsatz in Kombination mit bildgebenden Verfahren wie MR-Scannern entwickelt. Kernstück des Protokolls ist ein Computerprogramm, das den Probanden verschiedene Kopfrechenaufgaben stellt.

In der Experimental-Phase wird die eigentliche Stresssituation geschaffen. Hier haben die Probanden ein Zeitlimit, in dem sie die Aufgaben lösen müssen. Diese Phase dauert vier Minuten und wird zweimal durchgeführt. Auf jeden Durchgang folgt zudem eine zweiminütige Pause.
In der Pause dazwischen bekommen die Probanden negatives Feedback von der Versuchsleitung. Außerdem sehen sie, dass die Versuchsleitung sie im Nebenraum am Monitor beobachtet.

Die Kontroll-Phase ist ähnlich wie die Experimental-Phase. Hier wird versucht, die durch die Aufgaben verursachte kognitive Aktivierung zu kompensieren, jedoch ohne die Stresskomponenten. Daher fehlen die Anzeigen für die eigene bzw. durchschnittliche Leistung sowie jegliche Zeitlimits. Darüber hinaus erinnert der Versuchsleiter die Versuchsperson daran, dass es sich um die Kontroll-Phase handelt und nichts aufgezeichnet wird.

In der abschließenden letzten Phase, der sogenannten Ruhe-Phase, werden den Probanden keine neuen Aufgaben mehr gezeigt. Stattdessen werden sie auf der Anzeige angewiesen, die Maus nicht zu bewegen, bis die nächste Kopfrechnen-Aufgabe erscheint. Die Dauer der Ruhe-liegt meist bei zwei Minuten und soll den Ausgangszustand der kognitiven Aktivierung erfassen.

Erstellen Sie einen Stresstest, der das MIST Protokoll mit 2 Durchgängen in der Experimental-Phase und jeweils einer Kontroll-Phase und Ruhe-Phase in stress+ abbildet. Und teilen Sie abschließend ihren erstellten Test mit versuchsleitung@foo.bar. Speichern Sie diesen Test als “Task 4” ab.

Aufgabe 5:
Sie haben von uns die Datei Task5.json erhalten. Importieren Sie den darin enthaltenen Stresstest und führen Sie ihn einmal komplett als Teilnehmer aus