Conversational Interface for Screening

Master Thesis in Information Science

Author:

Robin Håvik

Advisor:

Frode Guribye

May 2018

Abstract

There are many adults who lives with ADHD without getting a diagnosis. When being evaluated for ADHD the first step is often to complete what is called the Adult ADHD Self-Report Scale (ASRS). ASRS is a symptom-check questionnaire built by the World Health Organization for screening adults for symptoms of ADHD.

In the study presented in this thesis, a prototype for a chatbot has been designed in order to explore how the ASRS test could be designed to a conversational interface. Having the ASRS in a conversational interface, users can answer questions from the ASRS with a more open language and supply answers with information that may be of interest for domain experts.

The prototype was evaluated amongst users by conducting a comparative experiment with two objectives. To find out how the results from the conversational interface differed from the results from the paper-based modality, and to find out how the participants perceived the prototype. The results from the experiment revealed an indication that the result differences were of non-significant and that most participants preferred the conversational interface to the paper-based modality. The results support that chatbots can be a useful technological utility for screening in the domain of mental health.

Acknowledgements

At first, I would like to express my sincere gratitude and thank my advisor Frode Guribye for all advice, guidance and encouragement I have received throughout the process working with this thesis.

Secondly, I would like to thank the INTROMAT project group, for letting me be a part of the group for the completion of my master's degree. I'm thankful for all the feedback and positivity.

I would like to thank Eivind Flobak for his assistance which helped to formulate the conceptual idea for the study presented in this thesis.

I would like to give a huge thank you to all the superheroes on room 539 for the good times these two past semesters.

Also, I would like to express my gratitude to my fellow master students Aleksander Tonheim, Anette Drønen Sunde, Elisabeth Wiken, Fredrik Madsen and Yara Mathisen for their support and motivation.

I would like to thank Sigve Solvaag for his effort to proofread this thesis.

At last, I want to thank all of the participants who participated in the study for their time and effort.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

It is common to occasionally to be inattentive in a meeting or experience having impulsive thoughts or behaviour. However, when these kind of symptoms causes larger issues in daily life situations, it could constitute having the neurodevelopmental disorder Attention Deficit Hyperactivity Disorder (ADHD) (Helsedirektoratet, 2014).

In Norway it is estimated that 3-5 % of the children and adolescents have ADHD and that two thirds of them lives with the symptoms as adults, which constitutes 2,5 % of the adult population (Helsedirektoratet, 2014). There are also adults who lives with symptoms of ADHD without having received a diagnosis (Hevrøy, 2016). With no treatment or ways to cope with the symptoms, the symptoms could have a negative impact on the daily lives for the adults, for instance at school, work or in social settings.

For an adult to get an ADHD diagnosis, the adult must go through a thorough evaluation process with domain experts. Before the evaluation process starts, it is common for the adult to complete what is called the "Adult ADHD Self-Report Scale" (ASRS) (Kessler et al., 2005), a symptom-check questionnaire used to find potential indications of signs or symptoms of ADHD. This test is paper-based and is either fulfilled by the adult individually or as in a conversation with a domain expert (Helsedirektoratet, 2014).

Today, chatbots are an up and coming way to interact with computers. More and more businesses are making use of chatbots as they are available instantaneous at all times for users, for tasks like for instance customer service (McTear, Callejas, & Griol, 2016a). Chatbots also has been applied to use for health and mental health related tasks, where it has been conducted research on how they can be used for assistive purposes. While there are a few symptom-check chatbots which exists today, limited research has been conducted on how conversational interfaces could be designed for screening purposes in mental health.

This thesis presents a study that has been conducted where the objective has been to design the ASRS test into a conversational interface, so a person can interact with a chatbot in order to get an indication if there are signs or symptoms of ADHD.

The study is done as part of the INTROMAT project. INTROMAT, which stands for '*INtroducing personalized TReatment Of Mental health problems using Adaptive Technology*', is one of three projects which has received funding as an IKTPLUSS Lighthouse project from The Norwegian Research Council in 2016. INTROMAT received funding for five years to develop innovative digital solutions for prevention, treatment and follow-up for mental health problems. INTROMAT's vision is to improve public mental health with innovative ICT solutions (INTROMAT, 2017). The project has five different prioritized case studies. One of them is cognitive training for ADHD which aims to study, design and implement solutions aimed for adults with ADHD (Intromat, 2016). The research presented in this thesis falls under this case.

1.1 Motivation

The objective of this thesis originally started out as a broad idea of designing a digital assistant application for adults with ADHD, where the design was built around a conversational interface. Due to limited time for development and an unclear vision of what the digital assistant was going do, the research changed direction. The idea of making a chatbot remained, but instead of making a full-fledged assistant, the research aim changed to designing a chatbot for conversational screening. Symptoms-check tests are often designed around a n-point Likert scale. This gives clear and precise responses to questions, but there may be questions around symptoms where it may be more challenging for a respondent to give a simple frequency-based response. By having it in a conversational interface, the idea was that a respondent could complete a symptom-check test and supply the answers with more contextual information that may be of a guiding character for a domain expert. As the prevalence of conversational interfaces is rising, it also was of interest to explore how one could make use of the technology and design a screening test to a conversational interface.

1.2 Research Questions

Following the motivation to conduct the research there has been outlined three questions, the first question is constructive whereas the two other questions are empirical. The following overarching research question was outlined for the research:

RQ1: How can we design a conversational interface for the ASRS test?

Following the design of the prototype, an experiment was conducted in order to evaluate it. For this purpose, two additional sub-questions were outlined:

RQ2: Will the results of the ASRS test be the same with a conversational interface and with a paper-based modality?

RQ3: How does the participants experience the conversational interface?

1.3 Structure of the Thesis

This list presents the structure of the thesis

Chapter 1 Introduces the aims of the study along with its problem space and research questions.

Chapter 2 Presents relevant literature for this project and relevant work.

Chapter 3 Describes methods that has been used to conduct this study.

Chapter 4 Describes how the artefact was designed and developed.

Chapter 5 Describes how the artefact was evaluated and the results of the evaluation.

Chapter 6 Discussion of the results of the evaluation set up against the research questions.

Chapter 7 Concludes the thesis with a summary of the findings with study along with propositions for future work.

Chapter 2

Background and Related Studies

The chapter presents and gives insight in the background and the related studies that are relevant for this study. At first is Human-computer interaction presented as a field of research. Further the chapter gives an overview over ADHD and a description of the Adult ADHD Self-Report Scale. In the end the chapter, relevant work found after conducting a literature review is presented. The literature review had its focus on Human-computer interaction in relation with ADHD, the development of conversational interfaces and how conversational interfaces are used today in different domains.

2.1 Human-Computer Interaction

Human-Computer interaction (HCI) is a multidisciplinary research field which has a focus on how humans (users) interact with a computer. HCI gained traction as a field of research in the 1980's at the same time as the personal computer gained popularity among the public.

The personal computer made computer technology more accessible for the public by offering personal software and hardware in a smaller format. HCI as a research discipline has a focus on the design, evaluation and implementation of interactive systems. As HCI as a field of research initially had a focus on personal computers, it has over time expanded to cover the design of a wider range of topics and devices related to information and communication technology.

HCI as a research field has since its early days been through a development which has changed its methods and how HCI researchers approach their subject. Bødker (2015) describes the development of HCI by dividing it into three phases referred to as *waves of HCI*. Bødker (2015, p. 24) characterizes the first wave as being driven by cognitive science and human factors, whereas in the second wave the focus shifted to how groups could use software applications in work settings. In a previous article by Bødker (2006, p. 1), she describes the changes in the second wave as *"rigid guidelines, formal methods, and systematic testing were mostly abandoned for proactive methods such as a variety of*

participatory design workshops, prototyping and contextual inquiries". Lastly, the third wave broadens the focus and brings attention to topics which received less attention in the past such as context, culture and values, along with the role of the researcher (Bødker, 2015). Harrison et al (2007) have conducted a similar analysis on the development of HCI and they refer to the phases as "the three paradigms of HCI".

2.1.1 HCI Research as Problem-Solving

HCI as a research field borrows some of its ideas and disciplines from other research fields, such as computer science, cognitive science, engineering, and social sciences. Although, what defines HCI it is its aim to *"to solve goals in human use of computers"* (Oulasvirta & Hornbæk, 2016, p. 4957). The "identity" of the field of HCI has for a long time been under debate because of the combination of the diverse ideas from the different fields. In an essay by Oulasvirta and Hornbæk, they do contribute with a meta-scientific account of HCI, where they see HCI as problem-solving research of three paradigms: empirical, conceptual and constructive (2016).

In the essay they do define empirical research as *"creating or elaborating descriptions of real-world phenomena related to human use of computing"* (Oulasvirta & Hornbæk, 2016, p. 4958). By this they mean to explore a phenomena novel to HCI research, discover relevant factors to the phenomenon, and in the end measure and quantify their effects on something of interest (Oulasvirta & Hornbæk, 2016).

Conceptual research is defined by Oulasvirta & Hornbæk as work that explores and explains *"previously unconnected phenomena occurring in interaction"* (2016, p. 4958). This type of research aims at tackling conceptual problems by making theories, concepts, methods, principles and models (Oulasvirta & Hornbæk, 2016).

Lastly, the aim of constructive research is *"producing understanding about the construction of an interactive artefact for some purpose in human use of computing."* (Oulasvirta & Hornbæk, 2016, p. 4958). The goal of constructive research is not the construction itself, but instead to understand the process with its ideas and principles (Oulasvirta & Hornbæk, 2016). For instance, a detailed documentation of a design process of an artefact, to justify the decisions that has been made for the design.

With these paradigms established, they defined a research problem in HCI as *".. a stated lack of understanding about some phenomenon in human use of computing, or stated*

inability to construct interactive technology to address that phenomenon for desired ends" (Oulasvirta & Hornbæk, 2016, p. 4960).

In HCI problem-solving it is common that the paradigms which has been described is combined with each other in one way or another. For instance, by conducting constructive-empirical research one could design a suggestion for an novel interaction modality and afterwards contribute to the understanding of relevant phenomena (Oulasvirta & Hornbæk, 2016).

In terms of this study, a prototype has been constructed (see Chapter 4), a chatbot for screening ADHD symptoms. Further, an empirical user-experiment was conducted in order to compare the conversational interface to the traditional paper-based modality of the ASRS test (see Chapter 5).

2.1.2 HCI and Conversational Interfaces

A conversational interface refers to an interface where it is possible to interact with a computer using natural language. In the field of conversational interfaces, it is possible to distinguish interfaces from each other depending on the way one interacts with them and how they are designed. There are for instance chatbots where the chatbot interacts with an user by the means of text (McTear, Callejas, & Griol, 2016b), whereas voice user interfaces is designed around using the voice as the primary input (Porcheron, Fischer, Reeves, & Sharples, 2018).

In the tech industry there have in the recent years been an optimism towards conversational interfaces as a way to interact with computers (Følstad & Brandtzæg, 2017). According to Luger & Sellen (2016), as conversational interfaces though have become more prevalent, there has been designed many poorly interfaces which do not meet the actual desires and needs of the users. Følstad & Brandtzæg (2017) touches upon the same topic and say that are many challenges reveal themselves when designing conversational interfaces and that conversational interfaces has not received enough attention from HCI researches. They therefore do argue that HCI researchers should embrace Human-Chatbot interaction as an area of design and practice. Though, according to Følstad et al (2018) in a more recent paper, the interest among researchers to research and design chatbots have now grown.

As the objective the of study have been to develop a conversational interface for the ASRS screening test is it a proposed contribution to the field of HCI.

2.2 ADHD – Attention Deficit Hyperactivity Disorder

This section will give a brief introduction to ADHD, how it effects adults, and current available treatment options for the disorder.

Attention Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental disorder characterized by three core symptoms: inattention, impulsivity and hyperactivity. According to Helsedirektoratet (2014) it is estimated that 3-5 % of children and adolescents have symptoms of ADHD, and that two thirds of them lives on with the symptoms in adulthood, in which covers around 2,5 % of the adult population in Norway.

The symptoms of ADHD can be divided into three core groups, divided by the frequency of the symptoms. The first category covers symptoms of impulsivity and hyperactivity, the second inattention, while the third category is a combination of both. The third group is the most common one.

By having problems with inattention, it is common to have struggles with for instance paying attention to and to organise activities. For the ones who struggles with inattentiveness in their daily lives, it could often lead to that they appear not to be listening, they have problems following instruction and it is easy for them to be distracted. It could also be harder for them to focus on an activity, which further can lead to them straight up avoids challenging tasks which needs continuous attention (Helsedirektoratet, 2014).

The group of people who only struggles with symptoms of hyperactivity and impulsivity is the least common one. In this group, it is common that the person has challenges with impulsive thoughts or actions. It could make a person do actions without thinking of the consequences, for instance interrupting in a conversation or having issues with turn-taking. Hyperactivity is not as common for adults, but for the adults and children who are it can be experienced as having extra energy that must be released. In practise it could lead to inappropriate behaviour, for instance having problems being silent or seated in a gathering (Helsedirektoratet, 2014).

2.2.1 Adults with ADHD

The research of this study is a proposed contribution to the cognitive training case for ADHD, which is a case which aims to create digital assistive technologies for adults with ADHD (Intromat, 2016). This subsection will therefore give a brief description on ADHD in regards of adults with the disorder.

ADHD has commonly been associated as a disorder which causes problems for children and adolescents, and of that reason there has been conducted less research on it in regards of adults having the disorder (Brown, 2008). Newer research does however show that symptoms of ADHD can persist into adulthood (Biederman & Faraone, 2005). It is common that symptoms will show themselves in the childhood for a person, but some symptoms may become more visible later as a person matures as a teenager or a young adult, because of the reason that the person gets more responsibility over own life decisions (Brown, 2008). Barkley et al (2008) argues it may be hard to detect ADHD for adults since the symptoms may not be as visible as they are for children, and that an adult has learned to prevent situations where the symptoms of the disorder may become a problem.

It is most common for adults with ADHD to have problems with inattention, for instance in meetings or in social situation. While impulsivity could effect an adult in social settings by making the adult interrupt or disturb other people, or by using money irresponsible (ADHD Norge, 2016a). According to Sinfield (2018), do adults have less problems with hyperactivity, since as most adults has matured they have also created coping strategies to control these symptoms in order to satisfy social expectations.

2.2.2 Treatment of ADHD

There are no methods to cure ADHD today, but there are ways to reduce the symptoms. Common treatment options today are medications and cognitive behaviour therapy. What causes ADHD is a reduced level of dopamine in the brain. To keep it short, the brain uses dopamine to regulate the transactions of signals from one nerve cell to another (ADHD Norge, 2016a). Medication like for instance Ritalin aims to stabilize the dopamine level for the person with ADHD. It is documented that medications works for 75 % of the people who uses it (ADHD Norge, 2016b). The medications do not cure ADHD, but it reduces the symptoms. Unfortunately, for some the medications could unleash side-effects (Sonne, Marshall, Obel, Thomsen, & Grønbaek, 2016).

Cognitive behaviour therapy (CBT) is used as a supplement for medication, especially for children and adolescents. CBT is revolved around learning to set routines and trying to create better habits to better overcome the symptoms of ADHD. At this moment, it is according to Sonne, Marshall et al (2016) limited how much research that has been done on how one could use technology to help the persons with ADHD, despite it being a prevalent disorder.

Adults with ADHD are offered few treatment options for their problems, despite the problems they experience in their everyday lives. Medication do reduce the symptoms, but sometimes it also may lead to side effects (Sonne, Marshall, et al., 2016).

2.3 Medical Screening

As the prototype designed for this study is a screening chatbot, this section will establish what constitutes medical screening. Furthermore, the Adult ADHD Self-Report Scale has been presented, as it is the screening test which has been designed to a conversational interface for this research.

Medical screening refers to either an evaluation of a population by using a test , or to use a standardised procedure in order to find a medical or psychological sickness which have not yet been detected (Braut, 2018). An example of a method used for screening is a standardised questionnaire which aims to find signs of symptoms based on the answers from a patient. A test like this could either be done by a patient himself or it could be done as in a conversation with a domain expert. The aim of a screening test is not to give a final medical diagnosis, but rather give an indication if a person should be closer examined by domain experts. Many questionnaire tests are structured to have a person answer how often he experiences a symptom. As symptoms may be something that may be experienced as relative over time, it is common to have a person complete a screening test multiple times over a longer time period to see if there are changes to the result (Helsedirektoratet, 2014).

According to Braut (2018), there are some issues tied to screening. From a medical perspective, the tests or the research methods must have a satisfying grade of sensitivity and specificity. Sensitivity refers to the ability of a test to correctly identify if a person has a sickness, whereas specificity refers to the ability of a test to identify if a person does not have a sickness (Bu, Skutle, Dahl, Løvaas, & van de Glind, 2012). If these criteria not are satisfactory, the results of from a test will not have much of a value. The screening test must also be rigid in such a grade that there is a low chance for the tests giving a person a false positive result. If a screening test returns many false positives, it could lead to giving the test a low validity grade and unnecessary costs. Preventing false positives is important, as if a test gives a false positive it could lead to false results, over-diagnosing, and for a patient create a sense of unnecessary insecurity around the patient's health situation (Braut, 2018).

2.3.1 Adult Self-Report Scale for ADHD

The World Health Organization (WHO) in cooperation with scientists from Harvard Medical School and New York University School have developed a symptom-check test for screening adults for symptoms of ADHD. The test is called "Adult ADHD Self-Report Scale" (ASRS) (Kessler et al., 2005). It is a standardised questionnaire which consist of totally 18 questions, where each question is related to a symptom in the DSM (Silverstein et al., 2018). The way the test is structured must a respondent answer each question in the test with an alternative from a five-point Likert-scale, where the alternatives range from "never" to "very often". Estimated time to finish the test are 5 – 10 minutes.

The ASRS consists of two parts, where the first part consists of 6 questions and the second part consists of 12 questions. In the first part, four questions concern inattention and the two last questions concerns hyperactivity/impulsivity. While combining both parts are there in total nine questions concerning inattention and nine questions concerning hyperactivity/impulsivity. The short ASRS test have proven to be the most decisive (Kessler et al., 2005) and is used for screening (Helsedirektoratet, 2014).

The ASRS test, as other tests, is not a tool which is meant to diagnose people with the ADHD diagnosis. It is rather meant to be used a guiding tool which can give an indication if a person is showing signs or symptoms that are consistent with the ADHD diagnosis. The ASRS test is often used as the first step towards getting evaluated for the diagnosis. The questions in the test have been designed to create a dialog between a domain expert and a patient to make it easier to determine if a patient is showing enough symptoms for a diagnosis. An ADHD diagnosis can only be received after a thorough process with a domain expert, often an expert with a psychological background (Helsedirektoratet, 2014).

As mentioned must methods used for screening satisfy strict requirements for the method to be valid. The ASRS test, has the test questions been designed to satisfy the DSM-V criteria and the test has proven to have good validity as it have a high grade of sensitivity and specificity (Adler et al., 2006; Silverstein et al., 2018).

In another study conducted by Bu et al (2012), the validity of the ASRS test was evaluated amongst patients who had substance use disorder (SUD). It was presented in the study that the ASRS test was able to correctly identify 94 % patients who had ADHD. According to Bu et al. (2012), a third of SUD patients have ADHD, the ASRS test by having such a high

validity level does then make it easier to give SUD patients a more adjusted treatment for their problems.

The ASRS test is the most used screening tool for screening adults for ADHD (Kessler et al., 2005), but according to Helsedirektoratet (2014) the following tests also have been used to evaluate if an adult is showing signs or symptoms of ADHD:

- Wender Utah Rating Scale (WURS) for ADHD for adults.
- Brown Attention-Deficit Disorder Scales (Brown ADD Scales)
- "Behaviour Rating Inventory of Executive Function" (BRIEF).

The ASRS test has been used in the design of the prototype for this study, as the test is the most used for screening adults for ADHD symptoms. The structure of the test also makes it viable for designing it into a conversational interface.

2.4 Related Work

A literature review was conducted to get an overview over the literature and work that is relevant for this study. This section presents an assistive technologies design framework for ADHD, a brief history of the development of conversational interfaces and how conversational interfaces are used today in different domains. To showcase the usage of conversational interfaces, a few apps designed around a conversational interface are presented, apps which exists in the commercial domain, and in the health and mental-health domains.

ACM Library and Google Scholar were primarily used as search engines to conduct the search for relevant scientific literature.

2.4.1 Assistive Technology Design Framework for ADHD

There is a lack of assistive technologies for users with ADHD according to Sonne, Marshall et al (2016). They have therefore built an assistive technology design framework in order to help HCI researchers design assistive technologies for users with ADHD. The framework is built to give HCI researchers a direction by looking at the problem in a technological dimension and in a dimension, which highlights the challenges in the ADHD domain. They have looked at previous studies, ADHD research, and related assistive technologies, and with the knowledge they built the framework.

Sonne, Marshall et al (2016) outlined three design principles they propose one should follow when designing assistive technologies for users with ADHD. The guidelines are:

> **1. Provide Structure to Facilitate Activities:** *"Structure is beneficial for people with ADHD, as they are more likely to succeed in completing tasks if they occur in a predictable pattern"* (Sonne, Marshall, et al., 2016, p. 67)
> **2. Minimize Distractions***: "(...) it is beneficial to limit external distractions in order to prevent people with ADHD from losing attention"* (Sonne, Marshall, et al., 2016, p. 67).
> **3. Encourage Praise and Rewards:** *"Praising and rewarding a child or a teenager with ADHD is a core element in parent training as this promotes desired behaviours"* (Sonne, Marshall, et al., 2016, p. 67).

They had children and adolescents in mind when outlining the principles.

2.4.2 Development of Conversational Interfaces

In the recent years, it has become more and more prevalent to interact with computers through a conversational interface, but conversational interfaces are not something new as research have been conducted on the subject since the 1964 with ELIZA (McTear et al., 2016a). ELIZA is known as the first chatbot and is a simple chatbot compared to the current state of art. It was able to analyse the linguistics of the sentences it received, and by looking for patterns in the sentences it found out what to respond based on conditional rules. According to McTear et al. (2016b), modern developments in technology such as more powerful processing, artificial intelligence, and the rise of the semantic web, they combined have made it possible to build more sophisticated conversational interfaces. The advancements in AI and machine learning technology brought huge improvements in speech recognition accuracy, spoken language understanding and dialog management. Developments of semantic technologies have also enabled agents to access unstructured and structured data on the internet almost instantaneous (McTear et al., 2016b).

Conversational assistants have become more prevalent since Apple unveiled Siri for the iPhone. Siri was perceived as having a "virtual butler" in the phone. Other competitors have followed Apple and made their own conversational assistant, Google with the Google Assistant, Microsoft with Cortana, and Amazon with Alexa. Each of the assistants does tasks that are predefined and can answer to fixed number of automated queries. (Fischer & Lam 2016).

Chatbots has been rising in areas such as educations, information retrieval, business, and e-commerce (McTear et al., 2016a). Facebook and Microsoft in 2016 endorsed

conversational interfaces and with it they released bot-frameworks which simplified the process of building chatbots and deploying them to the public, for instance through Facebook Messenger, or Skype. This led to a rising number of businesses making their own chatbots, these could be automated online assistants that can support or even replace human-provided service (McTear et al., 2016a).

What makes chatbots attractive for the commercial market is that it available for customers instantaneous at all times, which is practical for instance for customer service. Two examples of businesses in the commercial market who uses chatbots are Domino's and Nordea. Dominos in some of its markets has a chatbot which a customer can interact with to order a pizza from the restaurant (Perez, 2017). Nordea also recently in 2017 released a chatbot assistant named "Nova", a customer service chatbot for its banking customers. A customer can interact with Nova in order to get answers around frequently asked questions on topics concerning for instance online banking or practical information around the saving accounts the bank offers (Nordea News, 2017).

The prototype designed in this study is a chatbot where the input is text. It has been developed by using the service from IBM named Watson Assistant (see Chapter 4).
The further sections will present examples of chatbots that has been developed for the health and mental health domain.

2.4.3 ADA – The AI Doctor

In the healthcare domain there are some instances of chatbots powered by artificial intelligence that are supposed to resemble an "AI-doctor" which is available to patients at all times to respond to health-related questions. An example is ADA, the personal health companion made by the British and German startup named ADA (ADA, n.d.).

According to the founders of ADA in an interview with TechCrunch (2017a), users are able to interact with ADA by describing symptoms to it, it can so give information on what may be the cause of the symptoms and how one could treat them. ADA uses techniques from artificial intelligence and machine learning to learn and create a profile of the user based on the user's medical history, so ADA can give more personalised assistance. ADA is not designed to replace doctors, but it is rather a service which is designed to make it easier to make informed decision around health-related issues without having to involve a human doctor when it is not necessary. The founders behind ADA argued that by having users use

ADA for more trivial issues, doctors may be able to use their resources as efficient as possible.

ADA also does have a rival named Babylon Health, made by a UK startup (O'Hear, 2017b). Babylon Health has a similar AI symptom-check function as ADA, but a feature special for Babylon is that it makes it possible for users to get in touch with doctors and specialist through text and video (O'Hear, 2017b), whereas through ADA it is limited to text communication (O'Hear, 2017a).

2.4.4 Chatbot for Symptom Checking

Though there has been built some conversational assistants for symptom-checking in the commercial market there was little scientific literature found on the topic. Fisher & Lam (2016) have made a proof-of-concept for a chatbot for symptom-checking based on using a flow-chart from the *American Medical Association Family Medical Guide.* The book is a medical book aimed for non-medical people and the book has several flow charts that are supposed to help the reader to diagnose his problem by answering yes and no questions (Fischer & Lam 2016).

The design of the chatbot has been built around the flow chart that was mentioned, by doing this did Fisher & Lam (2016) limit what was possible for a user to respond to the bot by giving the user the option of answering yes or no. They describe the chatbot as being proactive, which means the chatbot steers the conversation and asks the questions in contrast to having the user ask the chatbot the questions. They argue for the benefits of having a proactive chatbot by saying that this would prevent a user asking questions that are out of the bot's domain, and secondly since the bot asks the questions it will then limit the topic of the conversation to what is relevant for the symptom checking.
The authors additionally built a crowd-sourcing framework which makes it possible to further train the chatbot with more data from the book that was previously mentioned.

2.4.5 Woebot – Chatbot for Cognitive Behaviour Therapy

Woebot is a chatbot that has been designed by scientists at Stanford University to deliver cognitive behaviour therapy by offering users short daily conversations and mood tracking (Fitzpatrick, Darcy, & Vierhile, 2017). In the conversation between a user and Woebot, the Woebot is the part who drives the conversation. Woebot asks users questions about how the user is feeling and what is going on user's life. The user has a set of predefined responses which are possible to use to respond to a question. The responses are tailored for

each question and they could be either a text or an emoji button, in which resembles the user's affection the closest. The bot's conversational style according to Fitzpatrick et al (2017, p. 3) has been designed around human clinical decision making and the dynamics of social discourse. Below are six aspects which has guided the design process when building Woebot:

> **Empathic response:** The bot is designed to respond to a user in an empathic way in which is appropriate to user's mood based on the given input.
> **Tailoring:** Specific content is sent to a user depending on the mood of the user. For instance, if the user experiences anxiety, the Woebot offers help that can guide the user through the event.
> **Goal setting:** Woebot asked the participants in the study about if they had a personal goal that they wished to obtain in the period of two weeks.
> **Accountability:** The bot sets expectations of regular check-ins and follow-ups to earlier activities in order to create a sense of accountability.
> **Motivation and engagement:** Woebot tried to engage the participants in the study by sending each user a personalized message daily in order to initiate a conversation. The chatbot used emojis and GIF's to encourage effort and completion of tasks.
> **Reflection:** Woebot provided the participants weekly charts which described the mood of the participant over time. All of the graphs that were sent to each participant were sent with a brief description in order to facilitate reflection.

Woebot originally was built for young adults in college and graduate school. In a study conducted at Stanford University, it was revealed that adults in the age between 18-28 years experienced reduced symptoms of anxiety and depression by using Woebot (Fitzpatrick et al., 2017). According to Fitzpatrick et al (2017), 85 % of the participants used Woebot daily or almost at a daily basis in the test period. From the results it were reported that those users found the conversational interface to be engaging and they also viewed Woebot more favourably than the information-only comparison (Fitzpatrick et al., 2017).

2.4.6 Embodied Conversational Agent for Healthcare

The previous conversational agents have been examples of chatbots where the conversation is presented in a text interface. There has also been done research on how embodied virtual agents (ECA) could be used in the healthcare domain. An embodied conversational agent is an agent which is embodied into an avatar. An example of this to embed the agent to a virtual human to enhance the interaction experience by simulating properties of face-to-face conversation, such as verbal and nonverbal behaviour (Provoost, Lau, Ruwaard, & Riper, 2017) .

SimSensei Kiosk (DeVault et al., 2014) is an example of such an agent. The SimSensei Kiosk is virtual human interviewer designed to create a more engaging face-to-face conversation in order to make the user feel more comfortable to talk and share information to the agent. The agent has been embodied in a virtual human named Ellie, who conducts semi-structured interviews. The interaction has been designed to make the interview sessions favourable to automatic assessment of psychological distress indicators, referring to verbal and nonverbal behaviour correlated with depression, anxiety, or post-traumatic stress disorder (PTSD) (DeVault et al., 2014).

According to DeVault et al. (2014), it was reported in the results from an evaluation among users that a majority of the participants were willing to share and felt comfortable sharing information revolving psychological distress to Ellie. Many of the participants did also share intimate information in the interaction. A minority of participants was on the other hand very happy with the agent's ability to sense the user's nonverbal behaviour.

2.5 Chapter Summary

This chapter has presented HCI as a research field and given a brief overview of the domain of conversational interface. Further, the chapter gave an introduction to ADHD and insight into screening and the Adult ADHD Self-Report Scale, the symptom check test which has played a crucial part in the development of ROB.

At last has the result from a literature review been presented to showcase related work for this research. Some chatbots from both the commercial market and the scientific community were presented, where some of them gave inspiration for the design and development of the prototype.

Chapter 3

Methodology

For this study was the following overarching research question outlined:

How can we design a conversational interface for the ASRS test?

This chapter presents methods and techniques that were applied to answer the defined research question. The methods and techniques presented in the chapter are presented to give insight in how they work and how they fit in the research design.

3.1 Design as Science

The complexity of the new systems have led to a need among researchers to have formalized procedures for design in relation to scientific research (Bayazit, 2004).

Thought about having a scientised design approach can be traced back to *De Stilj* in the 1920's (Bayazit, 2004). The idea was later actualized in the 1957 by Buckminister Fueller when he coined the term Design Science. Further, in 1962 the Conference of Design Methods were held in London, and the event resulted in giving design methodology a new status in the scientific community by making it a new subject of research (Cross, 2001).

The relationship between the topics of design and science have been thoroughly discussed in the scientific community (Cross, 2001). Design methodologists sought from early on to make a clear distinction between design and science.

> *The scientific method is a pattern of problem-solving behaviour employed in finding out the nature of what exists, whereas the design method is a pattern of behaviour employed in inventing things of value which do not yet exist. Science is analytic; design is constructive.* (Gregory, 1966, p. 6)

Design science at first did not consider an artefact as an important or proper source for knowledge contribution. As design science has been in development, so has the view on the artefact. In HCI research, there have been developed an approach where the hypothesis

of a research case is updated and re-framed repeatedly based on new knowledge that has been acquired by designing an artefact (Zimmerman, Forlizzi, & Evenson, 2007).

Cross (1982) in *Designerly Ways of Knowing* argues in favour of artefacts as a source of scientific knowledge. In the paper, he discusses how material objects in the past objects have been designed by observing existing objects in order to see what works in the current design, like shapes, sizes, and materials (Cross, 1982). By observing previous designs of object, one can learn and copy from what works in a design and discard what does not. He further argues that *"one does not have to understand mechanics, nor metallurgy, nor the molecular structure of timber, to know that an axe offers (or 'explains') a very effective way of splitting wood*" (Cross, 1982, p. 6). Cross (1982) believes an object can be a source of knowledge by observing how an object is designed and how it is used. With this he justifies the position of how scientifically designed artefacts are a viable source of scientific knowledge.

In the field of HCI the methodological framework named Research through design has been widely adopted by HCI researchers. Research through design as framework recognises artefacts that has been designed as a source of knowledge. The framework has been used in this research to structure the research and design processes to design an artefact and to get knowledge from the artefact that has been developed.

3.2 Research Through Design

Research Through Design (RtD) is a framework of design research proposed by Zimmerman et al (2007). In the proposed methodological framework Zimmerman et al (2007) do focus on how an interaction designer should work to create the "right thing" in HCI research, *"a product that transforms the world from its current state to a preferred state"* (Zimmerman et al., 2007, p. 493), in contrast to the industry where the focus lies on making commercially viable products.

Zimmerman et al (2007) had wicked problems in mind when they proposed the framework. A wicked problem is a problem that is vague or of such a complexity that it is hard to use traditional engineering methods to solve them. A wicked problem is initially a term which originates from organizational sciences, defined by Horst Rittel as: *"a class of social system problems, which are ill-formulated; where the information is confusing; where there are many clients and decision makers with conflicting values; and where the ramifications in the whole system are thoroughly confusing"* (Churchman, 1967, p. 1).

To handle these problems, they do propose in their framework that interaction designers should*:*

> *.. integrate the true knowledge (the models and theories from the behavioural scientist) with the how knowledge (the technical opportunities demonstrated by engineers). Design researchers ground their explorations in real knowledge produced by anthropologists and by design researchers performing the upfront research for a design project* (Zimmerman et al., 2007, p. 497).

A problem is a target for continuously iterative processes where potential solutions are invented and critiqued. The problem is re-framed continuously by design researchers in order to attempt to make the right thing. (Zimmerman et al., 2007)

3.2.1 Evaluation of the Design Process

To evaluate the design process of a research project Zimmerman et al (2007, p. 499) do provides four critical criteria which describes how to evaluate an artefact and to describe what constitutes as a good design research contribution for researchers that follows this framework. Below are the four criteria:

> **Process:** The process of how a research contribution is created is a critical aspect for judging its quality. Documenting the process makes it possible to examine the rigor of the methods that were used and why they were selected for the research project. Generally, in science it is a sign of high quality if it is possible to reproduce the result of a contribution. However, in HCI research similarly to other social sciences, it is not given that reproducing contribution will give the same results, but by documenting the process, the researchers must think through and give details on how an experiment should be conducted and why. This applies rigor to the research.
> **Invention:** It is critical that the contribution from a design research project offers something new to field if it going to be considered a contribution. Therefore, they argue that it is necessary to do a proper literature review in order to justify that the contribution offers something new to the research community.
> **Relevance:** As it was mentioned earlier, it is not expected that by reproducing a design research project that it will produce same results if it is done by another researcher. That is why in instead of applying validity as a criterion, should one instead look at relevance. They argue that designers should frame artefacts within the real world, and therefore researchers should describe what state the design of the artefact is trying to achieve and make an argument for why the scientific community should consider this to be the preferred state.
> **Extensibility:** The last criterion is extensibility. Extensibility means that a design research project should be described and documented in such a way that it is possible for other design researchers to use the results of a research contribution to

> *"either employing the process in a future design problem, or understanding and leveraging the knowledge created by the resulting artefacts"* (2007, p. 8).

Research through design has been used as the design research framework to structure the process of the research and to gain knowledge from the prototype that has been designed.

3.2.2 Why Research Through Design?

RTD was chosen as an overarching design research framework for this study due to it being a methodology tailored for HCI research and that the methodology acknowledges an artefact as a viable contribution to knowledge and research.

3.3 Prototyping

> *"Prototypes should command only as much time, effort, and investment as are needed to generate useful feedback and evolve an idea. The more "finished" a prototype seems, the less likely its creators will be to pay attention to and profit from feedback. The goal of prototyping isn't to finish. It is to learn about the strengths and weaknesses of the idea and to identify new directions that further prototypes might take"* (MacKenzie 2013, p. 128)

In HCI research and software development it is usual to make prototypes to see if an idea for a solution could work in order to solve a problem. According to Rogers Yvonne, Sharp Helen (2011), prototypes usually are distinguished into two separate categories, low-fidelity and high-fidelity prototypes. Low-fidelity prototypes are a way to visualize the design of an idea quickly and with few resources. Examples of low-fidelity prototypes could be design mock-ups, wireframes, and Wizard of Oz- demos. A low fidelity prototype does not represent a full-fledged implementation of an idea, since the interaction and functionality of such prototype is restricted. On the other hand, it does showcase in an uncomplicated way the vision one could have for an idea and how it could be designed.

While on the other hand, a high-fidelity prototype is a prototype which in terms of design and functionality is close to a finished concept. For a research through design project it is crucial to make a high-fidelity prototype to demonstrate the vision for what the right thing is.

3.4 Evaluation

In design research it is crucial to evaluate the prototype that have been created in order to find out if it is actually the "right thing" according to Zimmerman et al (2007). For the

evaluation of the prototype designed for this study, a controlled comparative experiment was conducted. The prototype was compared against the traditional paper-based ASRS test. This section will describe the methods used and shortly why they have been chosen for the research. The structure of the research experiment is presented in Chapter 5 (see subsection 5.1)

3.4.1 Controlled Experiment

The experimental method is a way of conducting research where the knowledge is acquired through controlled settings, for instance in a laboratory (MacKenzie, 2013). According to MacKenzie (2013, p. 130) knowledge may be acquired by studying new knowledge, but it can also be acquired by studying existing knowledge in order to verify, refute, correct, integrate, or extend that knowledge. Experiments conducted in a controlled setting will have less relevance, but more precision due to the tasks given are artificial and is done in a non-natural setting. On the other hand will a controlled experiment raise the precision of the data acquired by the fact that the influence from factors from the real world such diversity and chaos is reduced or removed entirely. (MacKenzie, 2013, p. 131).

To conduct a controlled experiment, it is necessary to have at least two variables: an independent variable and a dependent variable. In the context of HCI, an independent variable could be suggestions for an interface or an interaction technique. A dependent variable on the other hand is a property of human behaviour that is observable, quantifiable, and measurable (MacKenzie, 2013, p. 131). In other words, it is knowledge that can be acquired and compared when comparison of different designs is evaluated. A typical dependent variable is time, the time of completion to solve a task.

The experiment of the study has been conducted in a controlled setting and for the experiment there have been defined independent and dependent variables. As the experiment also was a comparative experiment will the next section describe what describes such an evaluation.

3.4.2 Comparative Evaluation

As the prototype of this study is a new interface for an already existing test, this makes it natural to compare the conversational interface with the paper-based ASRS test by conducting a comparative evaluation. This section will describe what describes a comparative evaluation.

According to Mackenzie (2013) evaluation in HCI research often does have a focus on analysing a single aspect, without comparing the aspect to others of similar character. He argues that more meaningful and insightful results are obtained if a comparative evaluation is conducted. In many cases by not comparing a new design or interaction with an alternative it will make it more challenging to determine if it is an improvement to the state of the art.

In practice, a comparative evaluation will take a suggestion for a new design or form of interaction and compare it with other alternatives. The alternatives could be suggestions to other new alternative design, an established design, or a combination of both. Comparing designs could give insight in performance, accuracy, ease of use, and give input from users on what they prefer after seeing different designs (MacKenzie 2013).

There has been conducted research on the viability of comparative research. In particular a study by Tohidi et al. (2006), the hypothesis of the study was that a comparative evaluation would yield more insight than a one-of evaluation, where only a single modality is evaluated. The study had participants who were split into separate groups, and they were supposed to manually perform simple tasks with climate control interfaces. Three interfaces were tested, and the study had some of the participants performing tasks on only one interface, while the other group tested all of them. The findings of the study revealed that the participants who tested all of the interfaces, they became more critical of the interfaces and became more observant to problems of the different designs when they had been exposed to them all (Tohidi et al., 2006).

3.4.2.1 Within-Subject Design

In HCI experiments when applying test conditions, it is common to use the model of within-subject design or between-subject design. The test conditions of this study are based on within-subject design. The test conditions being based on this model means that all the participants who are being evaluated in the study will be tested on all factors. Therefore, this model is also called repeated measures, since all the participants will do the same assignments. Using the between-subject design model would on the other hand mean that a participant would only be tested on one aspect.

According to Mackenzie (2013, p. 176) HCI researchers do prefer within-subjects design due to three specific advantages it offers over using between-subject design, those are:

1. It requires fewer participants, but that also means it requires more testing for each individual participant. Having fewer participants is less time consuming and requires less scheduling.
2. Secondly, the variance due to the participants predispositions will be about the same across the conditions in the evaluation. By predispositions in this context one refers to the aspects of the personality of the participants, conditions that may influence the performance in the test, for instance mental and physical condition. In practice this means if a participant is susceptible to be eager or tired that will also carry over across the different test assignments. In contrast, using between-subject design there must be more participants, which leads to a higher grade of variability because of the difference between each participant.
3. Lastly, it is not necessary to balance groups of participants, as there is a single group. In contrast to between-subject design in which has separate groups for each test assignment in the experiment. By using between-subject design, it is necessary to balance groups to ensure that the participants in the groups are equal when it comes to characteristics that may introduce bias that could influence the measurements of a test.

An implication of using the within-subject model is if a participant is tested on multiple factors, it could result in a learning effect. If there are two ways one could interact with something "A and B", if a participant first then interacts with A, it could influence how the participant interacts and experiences B.

The experiment presented in Chapter 5 has been designed after within-subject design, where the two modalities of the ASRS has been compared and the participants have tested both modalities.

3.4.3 Statistical Analysis

Will the results of the ASRS test be the same with a conversational interface and with a paper-based modality?

To answer the research question above, the answers to all the questions, from both modalities, they have been investigated by using the Chi-squared test.

The Chi-Squared test is a test commonly used for investigating the relationship of categorical data. The data is presented in a contingency table where the data is divided into categories and so does each cell in the table present the frequency of the observed data for

each category (MacKenzie, 2013). According to Lazar et al. (2017, p. 96), the chi-squared test has two assumptions that has to be in order for it to give a valid judgement. First, the data points in the table must be independent from each other, meaning that one participant can only contribute one data point in the contingency table. Secondly, the data samples should not be too small, and it is recommended that the total sample size have 20 observations or more.

In the user-experiment, it has been registered if there is a difference between the response from the paper-based modality and the conversational interface. The Chi-square test has been used in order to investigate if there is a statistical significant difference between the modalities.

3.4.4 Semi-structured Interview

How does the participants experience the conversational interface?

To answer this research question, it was determined to conduct semi-structured interviews with each participant recruited to the study, after the participant had finished the ASRS test in both modalities. A semi-structured interview is according to Rogers Yvonne, Sharp Ellen (2011), a type of interview which combine aspects of both structured and unstructured interviews, where the interview has both open and closed questions. The interview follows an interview guide which is similar for each interviewee, so that each person gets asked questions about the same topics. The interview starts with the questions from the script, and as the interview continues does the interview format open for follow-up questions where it is appropriate. The interviewee will be encouraged to talk till there are nothing more relevant to say about the given topic.

3.5 Chapter Summary

This chapter has presented the research design of the study presented in this thesis. To structure the design, and evaluation phases of the research a set of methods has been described. The methods and techniques presented has been used to answer the research questions of the study.

Chapter 4

Development of Prototype

This chapter covers the process of how ROB was designed and implemented. It describes how the requirements was established, the design choices were made, and technical documentation of the development.

In the end a total of three phases were completed to produce a high-fidelity prototype ROB – a chatbot which presents the Adult ADHD Self-Report Scale in a conversational interface.

Before going in detail on the content of each phase a summary is presented:

First phase: The requirements of the prototype were established. A conceptual design was made and at the end of this phase an early prototype had been developed. The flaws of the current prototype were explained.

Second phase: Measures have been done to improve the dialog experience. The prototype was presented to peers in INTROMAT at the end of this phase.

Third phase: To finish the development of the prototype an algorithm for result handling was written. Some design changes were also done to improve the usability of the prototype.

4.1 Tools for Development

This section presents an overview over tools and services used to make the prototype for this study. This includes the chatbot service that has been used and some utility tools which have provided security and structure to the development.

4.1.1 Watson Assistant

Watson Assistant[1] (formerly IBM Conversation) is a software as a service (SaaS) by IBM which aims to give developers cognitive tools to build conversational assistants for websites, applications, messaging platforms and IoT devices. The service is a fusion of two previous IBM services, 'IBM Conversation' and 'Watson Virtual Agent'. An assistant can be a broad term, so in terms of the Watson Assistant it refers to chatbots and voice agents. By combining the services, IBM aims to make it simpler for developers to build their own assistants that can be comparable to Alexa or Google Assistant (Vincent, 2018). Watson offers a set of tools for developing assistants, for instance tools for structuring dialogs, an API for natural language and tools for conversation analytics.

An instance of the Watson Assistant can be implemented into apps, websites, messaging services, as well as IoT devices.

4.1.2 GIT

GIT is an open source distributed version control system often used in software development. Version control is a type of software which observes, and controls changes of documents. Since GIT is a distributed system, it makes it easier for developers to work in teams by having local and distributed repositories. Using GIT over time will create a GIT "timeline" of the development process. If something wrong was to happen, there is a possibility to go back in the GIT timeline till a point where things worked as intended. GIT also supports creating alternative branches where it is possible to test and experiment functionality without having to effect the work in the main branch (Atlassian, n.d.).

GIT has been used in the development of this artefact to ensure that the code had version control and to have backup of all code in the development process.

[1] Watson Assistant - https://www.ibm.com/watson/ai-assistant/

4.1.3 Trello

Trello[2] is management tool for creating virtual boards in order to visualize task management (Trello, n.d.). Trello has been used to visualise the work board that has been used throughout the development of this artefact. In Trello it is possibly make columns with self-titled categories, and under each column one cat put up cards with tasks or user stories along where they fit in the workflow. The board in the development has been inspired by the traditional Kanban setup, where there are three columns “TODO”, “DOING”, and “DONE”. The board used in the development also has an additional extra column for tasks that were scheduled for the ongoing week.

4.1.4 NinjaMock

NinjaMock[3] is a tool used to make wireframes for applications and web pages (NinjaMock, n.d.). A wireframe is a schematic or a blueprint for how the design of an interface can look like. NinjaMock offers a simple interface with drag and drop and tools which makes it easy to sketch quick wireframes for a project. NinjaMock was used early in the development to illustrate possible designs for the prototype on a conceptual level.

4.2 Languages for Web Development

It early was decided in the design process to present the chatbot as a web application because of prior experience with web technologies. This application has been developed by using HTML5, CSS, and JavaScript at the frontend, while Node.js and IBM Cloud has been used on the backend. There was some consideration of using the JavaScript framework React, but as the web application only was supposed to show a chat interface it become clear that React would have made the development more complex than what would be strictly necessary, due to not having any prior experience with the framework. HTML, CSS, JavaScript and Node.js were therefore used for the development.

4.3 First Phase

This section describes the first phase of the development. The first phase had its focus on the conceptual design of the prototype and the early development of the prototype.

[2] Trello - https://trello.com
[3] NinjaMock - https://ninjamock.com

The development of the prototype began as the idea and the research question was defined. Before the development of the artefact started, several things were necessary to determine early in the process. The choice of what chatbot service to use for the development and what environment the chatbot should be presented in.

4.3.1 Choice of Technology

For the development of the artefact it was decided to use a bot-maker service to build a chatbot. By using a bot-maker service a lot of the internal logic behind the chatbot will be abstracted away from the developer and make it easier to focus on the chatbot itself. For the study project it was preferable not having to connect the chatbot to services like for instance Facebook Messenger. It was instead preferable to present the chatbot in an independent environment to preserve full control over the artefact and its data. Before landing on Watson Assistant some other services were also considered.

For the presentation of the chatbot it was a choice between developing a web application or an Android application. It landed quickly on developing a web application because it would make it possible for every device with a modern web browser to use the application. With a responsive interface, the chat interface would in practise be usable on mobiles, tablets, and computers. Also having more experience with web technologies, it made me feel more confident developing it as a web application.

4.3.1.1 Choice of Chatbot Service

As INTROMAT has a partnership with IBM, it made it possible for me to use Watson Assistant as a utility for the development. But before it was determined that I would use Watson Assistant, I also investigated for alternative chatbot services since Watson Assistant did not have support for Norwegian language.

There were several things that were considered before the decision landed on Watson. The criteria that guided the decision were;

- Tools and logic for structuring dialogs.
- Platform agnostic. It was preferable that the chatbot not was bound to a specific environment.
- Support for Norwegian language would have been preferable, but it was not required.

After finding several alternatives it ended up being a contest between Watson Assistant, Chatfuel[4], and Wit.ai[5]. Chatfuel and Wit.ai were free, while access to Watson Assistant was provided by INTROMAT. All services offered tools for managing and structuring dialogs, and two offered support for Norwegian language.

Chatfuel satisfied some of the criteria that were considered including having support for Norwegian language. Unfortunately, the chatbots made by using Chatfuel could only be deployed to Facebook Messenger. Being bound to Facebook Messenger was a deal breaker as the intention was to implement the chatbot into an independent environment, to have full control over the data. As it is a health chatbot there may be interchanged sensitive personal information, for that reason there was a wish to have full control over the application for this research.

Wit.ai is a service which recently was bought by Facebook. The service specializes in offering AI technology for natural language interpretation which can be used according to themselves for chatbots, mobile apps, home automation, wearable devices, and robots. Similarly, to Watson Assistant, Wit.ai had something called "Stories UI". This was a tool in the web interface of the service, which let developers make dialog trees to structure the flow of the dialog. Wit.ai also offered support for Norwegian language which made it appealing. Unfortunately, it was announced that the Stories UI feature was being phased out. The stories UI was planned to shut down in February 2018, so it would not be clever to depend on this feature when it was being phased out. Because of limited development time in the project there was a wish to use the time on the design of the dialog and its presentation instead of its underlying logic. Having no prior experience programming with Wit.ai or similar services it made me choose Watson Assistant.

4.3.1.2 Why Watson Assistant?

In the end it landed on using Watson Assistant for the development. It did not have support for Norwegian language, but as the objective for the study was to design a screening test to a conversational interface, it was not a problem of it being in English. Watson is not bound

[4]Chatfuel: https://chatfuel.com
[5]Wit.AI: https://wit.ai

to any messaging platforms and can be deployed to a broad range of devices and environments.

Further, Watson offered support for what is called intents and entities. Intents is in chatbot development phrases and sentences that one could couple to a category describing the intention of the users input. A command. Intents usually includes verbs and nouns. Let's say there is an intent for greetings in a conversation. Under such an intent one could collect a sample of phrases like "hello" and "good morning". By offering a set of examples Watson will be trained to understand that these and similar phrases are greetings, and then this intent could trigger Watson to greet the user. An intent is usually used a command to trigger a dialog sequence in a conversation.

Entities on the other hand could be described as parameter data in which supplies Watson with data that are relevant and describes the users input. Using entities is useful and makes it easier to create conversational conditions in a dialog which steers the direction of a conversation by modifying the intent depending on the content of the input. Watson Assistant has two types of entities (IBM, 2018):

> **System entities:** Those are common and pre-defined entities that by default can be used in all types of chatbots. These include numbers, e-mail addresses, dates, currencies etc.
>
> **User-created entities:** These types of entities are defined by developers of a chatbot and they resemble types of data that may be relevant for a specific type of chatbot. Examples of such types entities could be a list of types of animals, cars, or movies.

4.3.2 Establishing Requirements

A chatbot can be implemented in different ways, for instance can a conversation be based on writing in natural language, it can be driven by keywords, or buttons like seen with Woebot (see subsection 2.4.5).

All of the options that were mentioned has its upsides and downsides. Using primarily keywords and buttons will make it easier to design and use a chatbot if the use-case is to conduct a simple set of commands. Having only keywords or buttons will also limit and make it less open what can be said to a chatbot. Restricting what can be written to a chatbot can in some cases make it easier to use, due to how it for better or worse restricts the input options. This could prevent confusion on how to interact with the bot.

The chatbot in the end was determined to be text based with buttons for simple commands. That made the question of how was it to be implemented? As mentioned earlier are there many ways to implement a chatbot. As the conceptual idea of this chatbot was to screen adults for symptoms of ADHD, it made me adopt a central idea from the chatbot made by Fisher and Lam (see subsection 2.4.4). As it was argued in that research, it can be hard for a user to mention all symptoms that may be relevant and necessary to proceed and give a proper result for a symptom-check test. Therefore, is one of the design pillars of the design that the chatbot should be proactive and provide the structure to the dialog. By doing it in such a way it is necessary for the chatbot to have a set of relevant questions, as I do not have any expertise in this domain I have not created any questions. Instead the screening dialog has been designed by using the questions from the Adult ADHD Self-Report Scale (see subsection 2.3.1).

Another aspect important in the design of a chatbot is how it presents itself to the user. It is not uncommon to give a chatbot a type of personality to make its interaction with a user more engaging and satisfying. There are no fixed-answer on how the personality of chatbot should due to that there are different requirements and rules for what is appropriate in the context it is implemented for. As in the case of this chatbot there was a wish for it to present it in a professional and emphatic manner. The chatbot will receive sensitive data from a screening test that describes the behaviour and symptoms of a user. If then the personality of the chatbot is unprofessional, this could create a mistrust between the user and the chatbot. In which as a result could prevent the user from being sincere in his responses.

When these topics were explored there was in the end a research question and few requirements established. The development started with an overarching research questions which guided the process.

- *How can we design a conversational interface for the ASRS test?*

A screening test could itself be a module in which can be implemented into a digital assistant. There was some consideration of that the chatbot could possibly present information and advice about the ADHD diagnosis as a function on the side of the screening. Implementing an information aspect into the chatbot could be a useful thing on the side, but it would not have offered anything on the table in terms of answering the research questions of the study. It was considered to implement the functionality for the

prototype, but it was not prioritized. Having all this established it led to these outlined requirements;

- **R1:** The screening sequence should be based on the Adult ADHD Self-Report Scale (ASRS).
- **R2:** The structure of the ASRS test cannot be changed.
- **R3**: The presentation of the chatbot must be professional and emphatic.
- **R4:** The graphical user interface must be minimalistic with few distractions.
- **R5:** The user should be able to respond to the questions with natural language.

With the requirements set the development of the prototype could finally begin.

4.3.3 Conversation Structure

The development of the prototype began by designing a structure for the dialog in the Watson Assistant web interface. To build a dialog it as mentioned necessary to have intents, entities (see subsection 4.3.1.2) and a dialog flow which steers the conversation depending on the input. Four intents were outlined for this purpose:

- #greeting: Greets the user when user writes a greeting message.
- #goodbye: Says farewell and ends the conversation when the intent is triggered.
- #screening: Will start the screening sequence if triggered.
- #information: An intent that would be triggered if the user asked about ADHD.

Each intent has a set of phrases and sentences that the chatbot will react to. If an intent in Watson has been fed and trained with enough examples it could then also react to other sentences that are similar, but not explicit defined. As mentioned it was considered to make the prototype answer questions about ADHD, a general intent for this commando was therefore defined, but the information dialog was never properly implemented for the prototype.

By using the ASRS test as the design fundament for the prototype, there was a criterion which guided the construction of the dialog. The criterion, the structure of the ASRS test could not be changed, as it would void the validity of the result coming from the test. It was favourable to keep the result as valid as possible, this created a design challenge in the design process of the chatbot. The questions of the test could not be changed or rewritten, not the order of the questions, nor what is possible to respond to each question. This challenge is later addressed in the development.

The dialog flow in Watson Assistant follows an if/else tree structure to steer the conversation. Each dialog node in a conversation tree usually has a condition which must be satisfied for the node to be triggered. A condition can for instance be that it finds a certain intent or entity in the given input, or a more sophisticated condition where a variable is compared against a specific value.

The dialog in the prototype begins with having the chatbot named 'ROB' introduce himself and tell what he can do for the user. From there the user can apply to start the screening sequence by writing something which triggers the #screening intent. This could for instance be:

"I want to take the ASRS test"

By triggering this intent, the user will start the screening part of the dialog. Before ROB starts asking questions, the user is provided with some information about the test. It is established that ROB may not understand everything and that the user should respond how often he experiences the given symptom from the questions. If everything is fine, then the user must confirm if he wants to start the screening.

If the user confirms to start interaction, ROB will then ask the first question from ASRS test. In order to proceed to the next question, the user must respond with an answer which includes one of the five responses from the Likert-scale in the ASRS. Those which are 'never', 'rarely', 'sometimes', 'often', and 'very often'. These keywords have all been collected under a mutual entity named "@responses".

Figure 4.1 below presents a visualization on how the logic of the screening dialog works in practise.

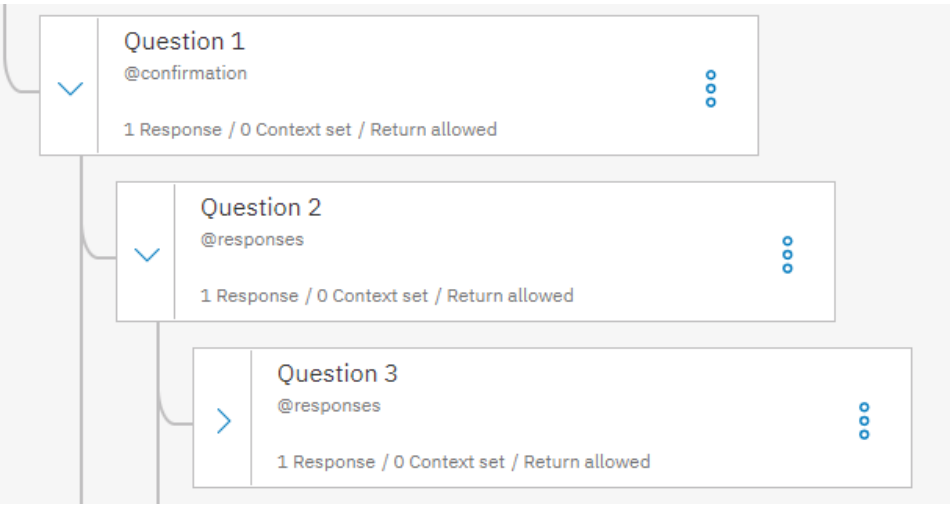


Figure 4.1 Snapshot of the dialog tree in Watson Assistant.

Only Part A of the ASRS were included in this phase, meaning the six first questions. At this point when the user had answered all six questions, the user would simply receive a thank you message, ending the conversation. At this point the user did not receive a result message with a score after responding to all questions. To calculate a result, one can assign a number to each of the response alternatives and see if the score reaches a certain threshold. Though, because of limitations in Watson Assistant it is not possible to perform arithmetic operations in the web application. Logic must be solved in the application in which the Watson Assistant has been implemented to. Meaning the local web application required an algorithm for this task. Having a dialog structure in place made it possible to begin the development of the web application.

4.3.4 Design and Implementation of Web Application

Before the development of the web application a few wireframe mock-ups were designed in Ninjamock. By using this tool, the wireframes quickly were sketched to give a conceptual view of how the application could look. The wireframes were made early in the process and some of the aspects presented in them may not be resembled in the final application. Two wireframes were made, one for a text-only interface, one for a button interface before it was determined that text was supposed to be the input. The wireframes were simple as it was a chat interface that was designed. One could argue on how necessary it was to make wireframes for such a simple interface, but having different wireframes to

showcase could be positive, due to how it makes it easier to showcase the different ways of interaction to users.

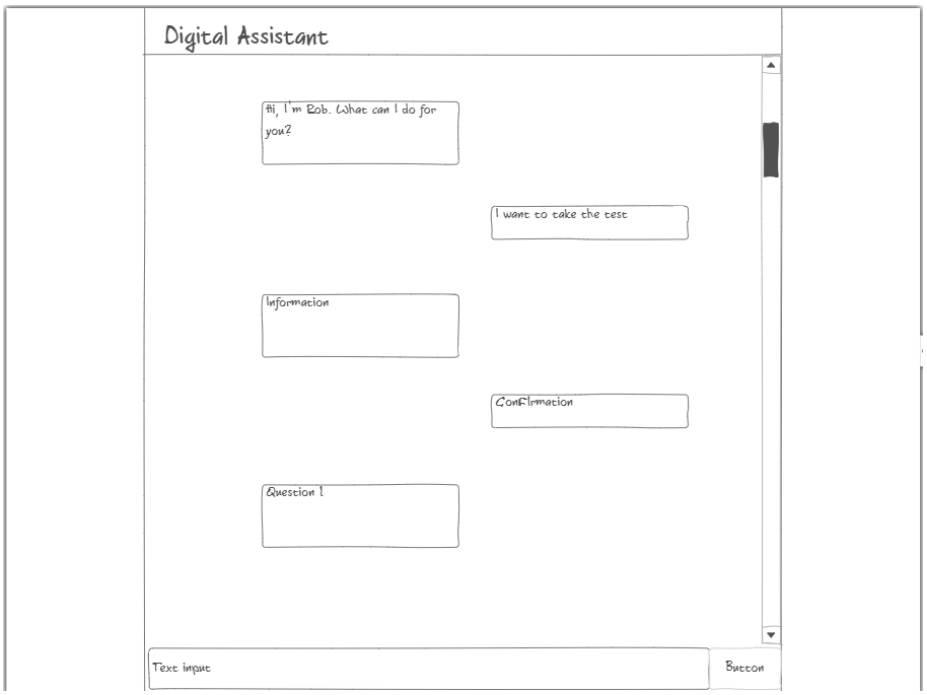


Figure 4.2 Design wireframe for the web interface

The following wireframe shows the chat interface as it was intended at a conceptual level. The intention was to keep the design minimalistic with a design with few distraction, and that the design may resemble an interface that are known from other messaging services.

The development of the web application was simplified to a degree when I stumbled across what is called 'Watson Assistant Sample Application". A web application in which implements the Watson Assistant with a simple and responsive interface. The sample app from IBM not only offers an interface, but also the internal logic behind in which handles the communication between the application and the backend in the IBM Cloud. The sample app is open source, which makes it viable to use it as a basis for extension for my own application. The app template as one may call it, can easily be forked from GitHub[6]. Having no prior experience with Watson API it really simplified the development by giving me more time to work on the dialog and its presentation rather than complex underlying logic necessary to make the application work as intended.

[6]Watson Assistant Sample Application: https://github.com/watson-developer-cloud/assistant-simple

A few design changes were done to web application to make it look better. When a user sends a message then the message is wrapped around a ‘bubble’, as seen in familiar messaging services. The colour of these bubbles was changed from having a light green colour to a flat blue colour. As blue is a calming colour, making it appropriate for a health application. A header for the webpage were also added for cosmetic purposes to give the page a title and remove some empty space at the top.

4.3.5 Result of First Phase

The first phase lasted for four weeks. In this phase, a conversational structure had been designed in the Watson Assistant web interface. A web application also had been developed so the app could be presented in the web browser. There were though still things that had to be done. For instance, if a user answered all the questions, the user would not get a result from ROB, only a thank you note. This is one of the prominent features that had to be addressed in the future phases. The dialog structure was not set in stone, so some changes in the structure and in the dialog were planned for further phases to make it feel and look better from a chatbot UX perspective.

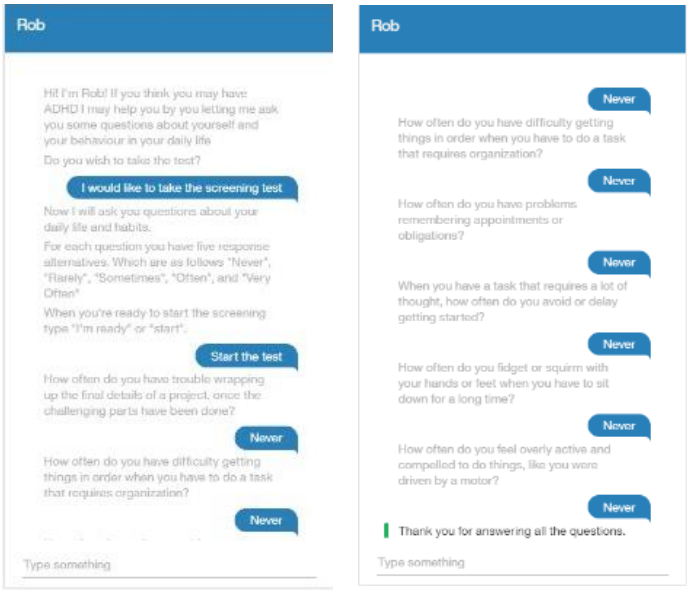


Figure 4.3 A & B Screenshots of the ROB in the mobile interface.

Figure 4.3 A & B shows how the prototype looked in its current form. The interface is simple and functional. There are though some problems with this implementation. In this implementation it is necessary for a user to use one of words from the Likert-scale in the ASRS test to proceed in the screening test. This makes it necessary for the user to either remember all the responses from the Likert-scale or make the ROB show the response alternatives in the conversation if there are problems. This made the conversation to

commando based and not so natural as it was intended. If a user is dependent on using the Likert-scale alternatives, it will most likely not motivate the user to give reflective responses. The implementation in its current form does not really differ from a traditional survey where the response alternatives is directly presented to the user, only difference is it may be harder to complete the form in this conversational interface.

Therefore, it became necessary to do some adjustments to face this challenge. The ASRS test is a standardised test, so the validity of the test would have been inflicted if the questions or the response alternatives are changed. Finding out how to make dialog more natural without breaking the validity was a design challenge in the design process. How the challenge was addressed is a topic in the second phase of the development.

4.4 Second phase

The second phase lasted for four weeks and in huge parts had its focus on the task to make the screening dialog more open, so users could give responses without being bound to having one of five keywords in their sentences. Measures were also done to improve the dialog structure.

4.4.1 Fallback Messages

It was explained earlier that after ROB has asked a question he will look for a response entity in the input from the user. Though if ROB did not find the entity it looked for, it would fall out of the conversation sequence. This was a problem, but it was not something that was complicated to fix. To prevent ROB from breaking the conversation thread a set of exception handlers were made, or what is called a "fallback message". A fallback message refers to an error message where the chatbot will tell the user that it did not understand the input (Barkin, 2016). It is not unusual to include tips for how to interact with a chatbot in a fallback message.

Fallback messages quickly were set-up and one were tied to each question node. If none of the entity-types were found in the input from the user, the fallback message would then be triggered. The message from the fallback tells the user that ROB did not understand the input. ROB would also in the same message give the user a message with a reminder on what user can write in the screening dialog. The following error message were outlined:

"I'm sorry, I didn't get that. Please answer the questions in a degree of never, rarely, sometimes, often, and very often"

The user is also able to trigger this fallback by simply asking ROB for a reminder of what is possible to respond to a question. After the fallback message has been sent, ROB will re-ask the same question the user was unable to respond to.

4.4.2 Synonyms for Enhancement of the Dialog

Changes had to be done to make the dialog between the user and ROB feel more natural. In its current implementation, ROB only understood the input if one of five words were found in the input of the user. To solve this problem an intuitive solution was found, applying frequency-based synonyms to each of the entity-types.

In the Watson web interface, it is possible to attach synonyms to entity types. A synonym is a word which has the same meaning or closely resembles another word. In practise this means that ROB still would look for the defined entities in the input from the user, but it would also automatically check to see if any of the words matches with one of the defined synonyms.

The synonyms used for ROB have been retrieved from thesaurus[7]. A thesaurus is an encyclopaedia for synonyms. It provided a set of synonyms for all the different response alternatives used in ASRS test. Attaching synonyms to the entities made it possible for users to write more naturally to ROB without being bound to using a few selected keywords, which is preferable in a conversational interface. Another thought for using synonyms, not only does it loosen up the rules of the conversation, but it may also keep the validity of the test in check. As the synonyms closely resembles or has the identical meaning to the words in the original test. The image below shows a set of synonyms attached to an entity in the Watson web app.

[7] TheSaurus: http://www.thesaurus.com

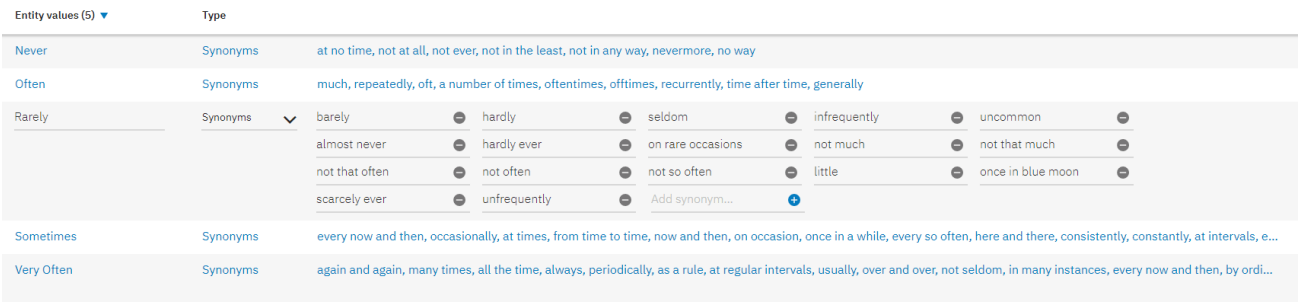


Figure 4.4 Screenshot showing the synonym overview page in Watson Assistant.

Using synonyms is a simple solution to loosen up the dialog, but it does also create new challenges for around the design. For instance, the synonyms to “often” and “very often”, they do overlap to a certain degree. This made it harder to apply those synonyms, since I had to do consider word for word and put them in the ‘category’ where they fit the best. In some cases, there may also be words or phrases which has not been defined as synonyms in which ROB does not detect. Another problem may be that a user could be using negations of a word, this may trigger one value while it is the opposite which is intended. These cases are harder to predict.

If a user also was to trigger several of the keywords because of the vast range of synonyms, this could make the chatbot pick the wrong synonym. In the current implementation of ROB, it only proceeded with the first entity value it found. It is hard to make such an implementation perfect, but in the case of this study it is satisfactory for inspecting and get a view of how a conversational interface for the ASRS could be designed and how it is perceived by users.

4.4.3 Feedback from INTROMAT

At the end of this phase the current prototype was presented in an INTROMAT meeting for the ADHD case. The meeting consisted of domain experts from different backgrounds, and also a user panel with two individuals with an ADHD diagnosis.

In the meeting I talked about the idea of making this chatbot. I showcased some of the wireframes that I had made before the implementation of the chatbot. In the end ROB was presented.

The feedback for the prototype was positive. One user in the panel argued that this looked like a better way on how one could conduct a screening test. A more engaging way in comparison to conducting the test in a traditional schema. The user also had some thoughts on how it could be better. Referring to one of the wireframes, the user thought more buttons

could potentially make the chatbot easier to use. It was also mentioned that it is not unusual that people with ADHD may also have dyslexia. Therefore, it could be an improvement if text-to-speech were implemented to ROB. Having this could make it easier for someone with writing problems to communicate with the chatbot by using voice input.

4.4.4 Result of the Phase

At the end of phase ROB had become slightly improved. The noteworthy achievements of this phase were the implementation of fallback messages and the synonyms. Both achievements enhanced the dialog experience. Attaching synonyms made it easier to interact with ROB, as the user not was strictly to the pre-defined words from the Likert-scale. Having fallback messages is kind of a security measure as it prevents the dialog from breaking. Also, does it give some additional information to the user to make it easier to continue from where the fallback occurred.

4.5 Third Phase

The third and last phase lasted for five weeks. During this phase the last touches on the application were done. There was a focus on two things ROB lacked a way to process the result it received from the responses from the user. To get this in place, it was necessary to implement an algorithm in the local app for handling this task. The design and the dialog in the application did also require some polish in order to make it ready for the user experiment. Some of the messages therefore were rewritten and the second part of the ASRS test was implemented into the dialog.

4.5.1 Result Algorithm

As mentioned earlier Watson Assistant did not have support for performing arithmetic operations, therefore it became necessary to do this locally in the web application. As it was a web application, a script was written in JavaScript.

It took some time to understand how to fetch and manipulate data from the Watson Assistant. To understand the logic, there are a few concepts that needs to be established and explained. The first one is what is called a context variable. It refers to a variable in Watson Assistant in which can store a value from either the input from a user or from an outer source. A context variable depending on its state can be used to steer a dialog. The role the context variable plays will be addressed as the other aspects has been established.

Another thing which is important to know is that all nodes in the Watson Assistant, and all messages are JSON-objects. JSON (JavaScript Object Notation) is a light data-interchange format based on JavaScript (Mozilla, n.d.). The web application sends and receives a JSON-object from the IBM Cloud, in which preserves the state of the dialog. Watson Assistant is stateless without it, so it is necessary to have this mechanism in order to have a dialog. The JSON-objects stores certain attributes, like for instance intents or entities the message has triggered. To process the responses from the test it is necessary to know which entity the user has triggered, this data has been extracted from the JSON-objects sent from Watson.

When the data has been received it needs to be processed. It is necessary to check if the message object stores a @responses entity, the entity which stores the response value that has triggered ROB. If the script finds said entity, it then will proceed to check the type of the entity. The function written for this task checks the type of the entity and depending on its type it will push a number into an array collection. For instance, if the user has triggered the “never” entity then the script will push a number of 0 into the array, or if it is “very often” it will will be a number of 4.

```
function responseToNumber(payload) {

        var response = payload.entities[0].value;

        if (response === 'Never') {
            result.push(0);
        } else if (response === 'Rarely') {
            result.push(1);
        } else if (response === 'Sometimes') {
            result.push(2);
        } else if (response === 'Often') {
            result.push(3);
        } else if (response === 'Very Often') {
            result.push(4);
        }
    }
```

Figure 4.5 Code snippet showing how a response gets converted to a number.

The script checks the size of the array and it continues to fill it up until it has six elements. When it has received six elements, it means the user has finished the first part of ASRS test. This will set in motion the second part of the script and it is here the context variable comes into the picture.

The node before the result node in the Watson conversation tree holds a Boolean context variable, which as default is set as false. The algorithm in the app calculates the numbers stored in the array if it has six elements. If the sum of the calculation reaches a certain threshold, the script will then return a value of “true”. This value will further be pushed up to Watson into the mentioned node and change the variable to true. The result node will then give a response according to this said value. If it is true, ROB will then return a message which says that there is a probability that the user may have ADHD, or the opposite if the value is false.

After the user has received the result after completing the first part of the ASRS test, the user will be asked by ROB to respond to some further questions. The user does not receive a result after answering the questions in the second part due to the nature of the ASRS test. The second part of the test is as mentioned used to further describe personal behaviour, so the data can be used as a supplement to result provided by Part A. The user receives a thank you message after answering all the questions.

4.5.2 Refinement of Design

When the result mechanism was implemented, it led the way to refine the existing aspects of the application in order to make it more polished. The refinement of the design can be boiled down to two things. Improvement of the dialog script and the inclusion of buttons. When writing the script for ROB, the writing was guided by some thoughts. A goal was to write the messages to be short and concise. If a message from the chatbot was long, it was then broken apart into separate messages. This makes the task of reading easier for the user. In the redesign of the dialog this was something that were guiding the writing process. It should also not be challenging for a user to find out how he interacts with a bot. To streamline the structure and interaction process a bit a few buttons were included for simple commands in the dialog, for instance for initiating the screening sequence. Buttons were not included in the screening test itself. The fallback response in the screening also was rewritten in order to not give an instruction that could influence the result too much. From listing up the response alternatives it now asked the user to respond to a question in a grade of how much a symptom occurred.

4.5.3 The Final Prototype

The end of the third phase marked the end of the development of the prototype. A fully functional high-fidelity prototype had been developed and it was ready to be evaluated.

After this phase, ROB now had the ability to calculate a screening result, a quite crucial feature that needed to be in place for the upcoming user experiment. The dialog had been polished and some adjustments had been done to improve the user experience.

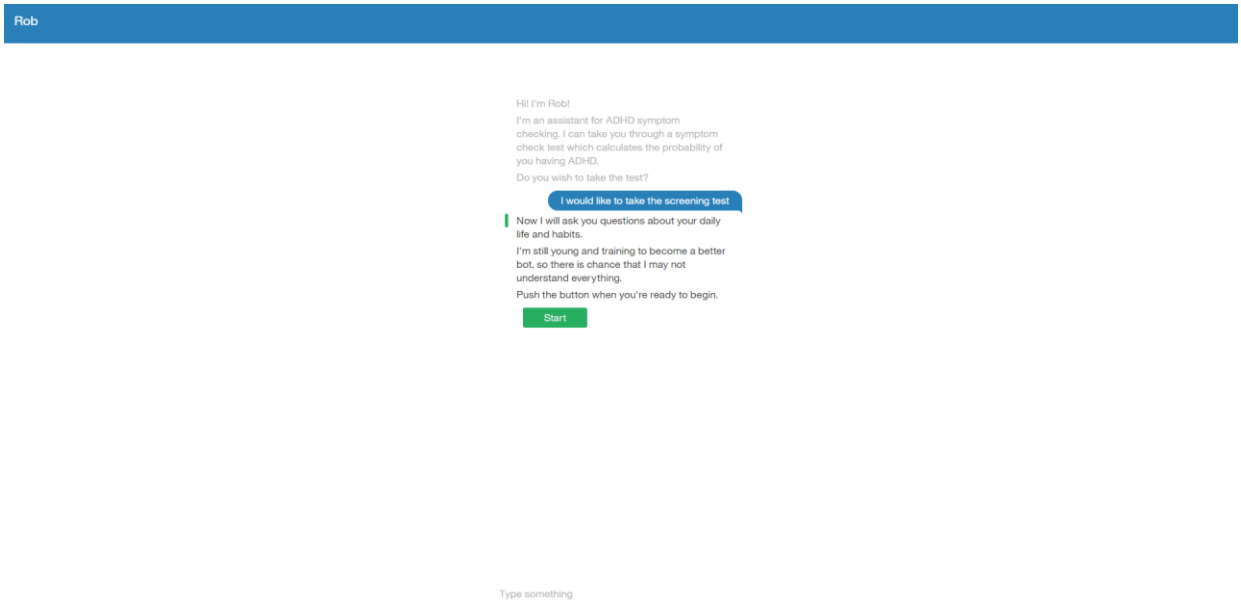


Figure 4.6 Screenshot of ROB in the PC interface.

Figure 4.6 shows the final prototype as presented in a web browser on a PC. Inspired by Woebot (see subsection 2.4.4), buttons were added to the dialog to conduct simple commands, as the upper image shows. The thought of having buttons for such commands is to remove uncertainty around how to proceed in the conversation in the areas of the dialog where there are buttons.

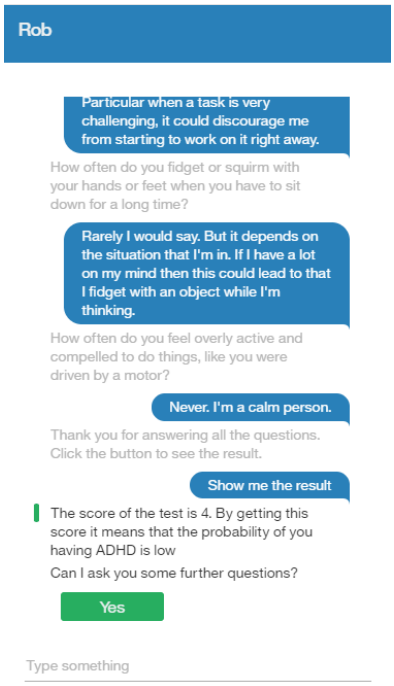


Figure 4.7 Screenshot with reflective responses and a result score.

Figure 4.7 shows the application in the mobile interface for better readability. The figure shows how a screening could look like by giving the user the ability to write and reflect around the questions given in the test. The figure also shows the user getting a score from ROB, and the user is also asked if he wants to answer some further questions, something which leads to Part B. There is only a "Yes" button included in this prototype in order to encourage test users to complete the whole ASRS test for evaluation purposes.

4.5.4 Discarded Features

There were some features that were discarded because of technical issues and time restraints. The features mentioned here is features that likely would have been implemented in ideal future version of ROB.

It is possible for a user in a conversation to trigger multiple entities in a response when giving an open answer. ROB unfortunately does not have the ability to choose the entity which stands the user the closest. There was an idea to make ROB ask the user to specify the entity which stood the user the closest if the situation were to happen. This feature was discarded due to uncertainty of implementation. ROB does instead proceed with the first entity it finds.

In the INTROMAT meeting referred to in the second phase there came a request of implementing a speech-to-text function in the prototype. It was considered, but it was

discarded as it did not provide enough value to prototype regarding the focus of the study. The study had a focus on written input, and as it were primarily to be tested with students from the Institute of Information and Media Studies. It was therefore discarded.

4.6 Chapter Summary

This chapter described the process of developing ROB through three separate development phases. The purpose of the chapter was to describe how ROB has been developed and to explain the design choices that has been made in the process. The result of the development was a functional high-fidelity prototype for an ADHD screening chatbot. ROB lets user respond to questions in the Adult ADHD Self-Report Scale with open language.

The next chapter will present the evaluation of ROB that was conducted in the form of a comparative experiment

Chapter 5

Evaluation

Having developed a functional prototype of ROB – a chatbot which presents the Adult ADHD Self-Report Scale in a conversational interface, it was now necessary to evaluate the prototype. A comparative experiment was conducted, where the objective was to compare the conversational interface to the traditional paper-based ASRS test.

The chapter describes the experiment and presents the results.

5.1 The Experiment

As ROB is a chatbot which lets a user complete the ARSR test in a conversational interface, the case of the evaluation became to conduct a comparative evaluation of the conversational interface and the traditional paper-based schema modality.

To conduct the experiment, 11 participants had been recruited, where the majority of the participants were master students from the institute, while a single participant was a bachelor student from another faculty. As it was a controlled comparative experiment, it was set up in a within-subject design model, meaning that each participant would in the experiment complete the ASRS test in the conversational interface and in the paper-based schema version of the test.

Having prerequisite knowledge of the ASRS test may cause a learning effect as it was mentioned in subsection 3.4.2.1). Therefore, to study this learning effect, the participant population was split into two sub-groups, controlling the order of what modality the participant was supposed to be exposed to first. Chatbot/Schema for the first group and Schema/Chatbot for the second group. A motivating factor of the designing the evaluation in this was to find out if the learning effect could have an influence on the dependent variables defined for the study.

Two independent variables and three dependent variables were defined for the experiment. The first independent variable is the modality of the ASRS test, the conversational interface

and the paper-based test. The second variable is if the participant has prerequisite knowledge of the ASRS test or not. Because prerequisite knowledge of the ASRS test may lead to a learning effect, in which can influence the dependent variables. When the participant had finished the tests, three dependent variables were then registered. They were, the result score from part A of the ASRS test, the participant's responses to the questions, and the time of completion.

The participants first were introduced to the ASRS test as a test used for mapping symptoms and calculating the probability of an adult having ADHD. When this was established the participants got an explanation about the experiment and what they were going to do. Most participants except for three had no prior knowledge about the ASRS test. Before starting each experiment did each participant sign a consent form, which gave me the permission to use the data collected from the experiment for the research. There was an emphasis in the consent form that no person was supposed to be able to be identified in the thesis.

The experiment was set up so that the users completed the test in a web browser on a PC and on a questionnaire on paper, where the user marked crosses for each response.

After each participant completed the test in both modalities had a set of participant data been collected. The responses from the participants, the result scores from part A of the test, and the time of completion for each test. There was also a debriefing after the experiment to get insight in the participant's experiences through a semi-structured interview. Each interview was recorded using a phone, and the recordings was later transcribed as close to the source as possible. Selected quotes from interviews have been translated to English and in the process of the translation was there taken care to preserve the participants intentions from the original transcripts.

By conducting this experiment, there was an objective to collect data around both quantitative and qualitative aspects. As ROB is an alternative implementation of the ASRS test, it was of interest to find out how the result of a screening test could potentially change. If the result differences between ROB and the schema are too large, it could void the validity of the test. For the study was it interesting to investigate the differences between the modalities regarding the results from the test and the responses to the questions.

On the qualitative side there was an interest in getting insight on what a participant wrote to ROB, how the user experienced ROB, and how it worked out having the ASRS in a conversational interface. There was a focus on getting the users to reflect around the

experiment by presenting their views around how it was use ROB in contrast with the schema solution. The participants were asked open questions to make them reflect around the questions asked in the interview.

5.1.1 Pilot Test

Before the user-experiment began it was conducted a pilot test to test the procedure that was planned. It was done to find out if there was something that did not work as intended, either with the procedure or with the prototype itself. The pilot test unveiled a bug in the prototype tied to the fallback mechanism, where if a fallback was activated in part B of the screening, the user would be sent forward to the next question instead of being re-asked the question the user was unable to respond to properly. As the bug was tied to Part B of the ASRS test did it not effect the result score received from part A of the ASRS-test.

The pilot test was a useful procedure as it unveiled the bug that was mentioned, and it did give me some experience of running the procedure before running it with the other 10 participants who were recruited to the study.

5.2 Analysis of Quantitative Data

Will the results of the ASRS test be the same with a conversational interface and with a paper-based modality?

In this section is the quantitative data from the user-experiment presented and analysed to answer this question. Before conducting the experiment, two hypotheses were outlined. A null-hypothesis and a regular hypothesis.

- H_0: The modality of interaction influences responses.
- H_1: The modality does not have an impact on the result.

To address the hypotheses, the answers to each of the questions in both modalities were compared and their relationships were investigated by using the Chi-square test.

Table 1 Results from user experiment

Participant	Chatbot	Schema	Time Chatbot	Time Schema	Error rate
0 (pilot)	12	9	05:30 min	03:00 min	3
1	19	12	04:30 min	02:45 min	7
2	14	8	06:22 min	01:56 min	6
3	10	9	03:12 min	01:45 min	1
4	6	4	03:29 min	01:41 min	2
5	12	11	10:08 min	02:10 min	1
6	3	3	02:20 min	02:30 min	0
7	12	9	09:22 min	04:10 min	3
8	7	8	02:37 min	02:06 min	-1
9	12	12	05:04 min	02:26 min	0
10	5	7	04:15 min	02:28 min	-2

Table 5.1 presents the quantitative data that has been collected from all the participants, including the pilot test. In total there are 11 participants, where each participant is distinguished by a number. The participants ranging from 0 – 5 belongs in group A, where they first were exposed to the conversational modality and then to the schema. The participants ranging from 6 – 10 has completed the tests in the opposite order.

For each participant, the table presents the result sum which has been retrieved from Part A from the ASRS test in each of the modalities. The time of completion also was recorded during the test to get a perspective on how quickly a participant completed the test in each modality. The last column presents the error rate, the response difference between two modalities.

The paper-based ASRS test is the most valid result from a default perceptive. Any difference from the paper-based modality will therefore be presented as an error rate.

By observing the data there are some patterns that are interesting to take notice to. In group A, the scores the participants have received from the conversational interface seems to be of a higher value than the result score from the schema. The average error rate of the group is of 3.33 points. The participants in the group did also seem to use more time completing the test, as most of them had no prerequisite knowledge of the test. The group had an average completion time of 05:32 minutes, while the average time for completing the schema was 02:12 minutes.

In group B, with the participants who completed the schema first, there were some other patterns. It seems like in this group being exposed to the schema test before taking the test in the conversational interface does have an influence the result. While in group A there was a tendency to that the result from the conversational interface was of a higher sum, in this group it is the opposite. In group B, the tendency was that the result from the conversational interface either had a lower score compared to the schema or not a difference at all. Only one participant in this group received a higher score in conversational interface. The average error rate of this is group is of 0 points. The participants also used less time to complete the test in the conversational interface and more time to complete the test in the schema when it was first presented. Making the time difference between the two groups, between the conversational interface and the schema of 49 seconds and 32 second respectively.

By looking at the numbers one can determine that result differences will present themselves depending on what modality one is exposed to first. A tendency which showed itself in group A, it was that the result they received from the chatbot was of a higher value than the one received from the schema, while in the other group it was the opposite.

Some of the participants received a high score in the conversational interface because of weaknesses in the design of ROB. Since this is a chatbot looking for certain words there were some issues regarding negations of words and phrases that has not been registered in ROB's thesaurus. For instance, participant #1 had an error rate of 7 points between the modalities. By reviewing the chat log, it was revealed that the participant had written "not very often" two times. This phrase was not explicitly registered in Watson, in which led to a misinterpretation in the result calculation. This response could be interpreted as "rarely" in the ASRS schema, but it was interpreted by ROB as "very often". This design flaw gave the participant a higher score than the participant should have had. This does exemplify one

of the weaknesses in the current design. It was later corrected to “rarely” in the corrected data.

Table 2 Presents the result of the experiment with proper representation of intent.

Participant	Chatbot	Schema	Error rate
#0 (pilot)	9	9	0
#1	13	12	1
#2	11	8	3
#3	10	9	1
#4	6	4	2
#5	12	11	1
#6	3	3	0
#7	9	9	0
#8	7	8	-1
#9	12	12	0
#10	5	7	-2

Table 5.1 did not present the result in a way which represented the intent of all participants properly. The conversation transcripts were therefore revisited to find responses that were misinterpreted by ROB. The misinterpreted responses were detected and carefully corrected so the intent of the participants was properly presented in the data. Table 5.2 presents the results in regards of how they would have been if ROB had not misinterpreted the responses from some of the participants. By cleaning up the data, the average error rates in the groups changed to 1.3 points and – 0.5 points respectively. The same data tendencies mentioned earlier did reveal themselves in the corrected data.

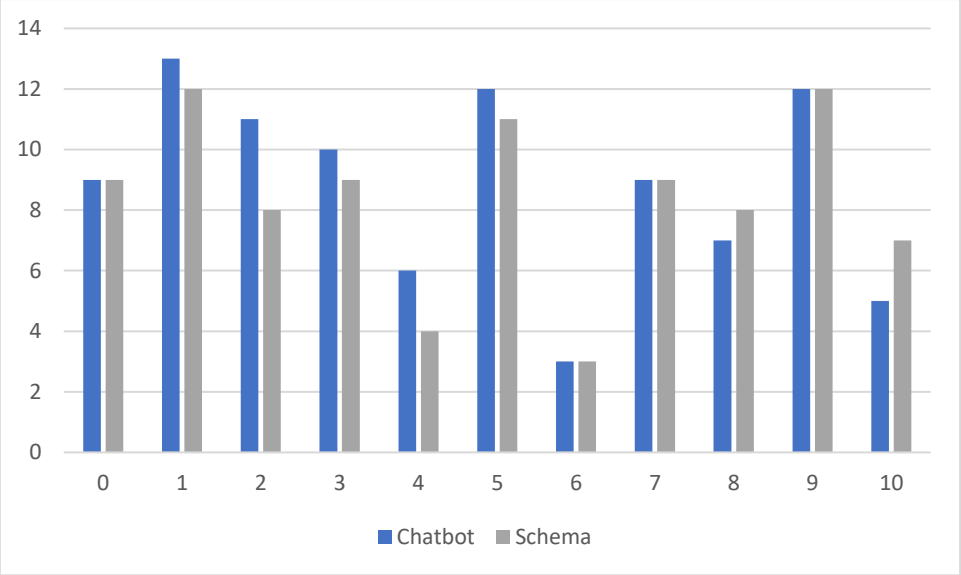


Figure 5.1 The results each participant received from the ASRS

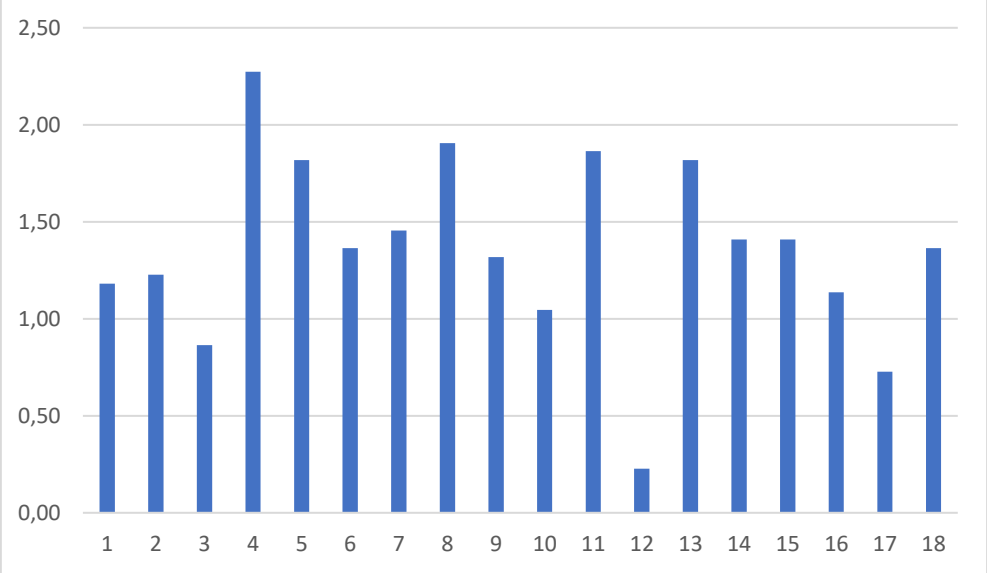


Figure 5.2 Average score for each question

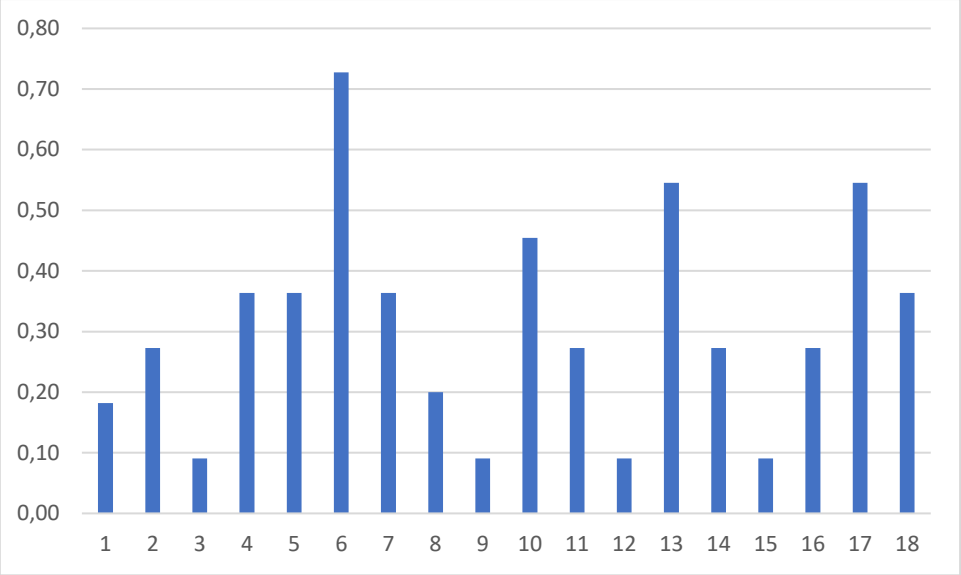


Figure 5.3 Average error rate for all questions

It was of interest to investigate what the participants had answered to each of the questions in both modalities. Figure 5.2 presents the average score for all the questions from the participants. Questions 4 had the highest average score of 2,27, while question 12 had the lowest average score of 0,23.

For tests like the ASRS it is common to take the test more than once in order to see if symptoms or results may change over time (Helsedirektoratet, 2014). Taking a test in one modality before the other may also cause a learning effect. The responses for all 18 questions were compared to find out if the participants answered differently to the questions from modality to modality. Figure 5.3 presents the average error rates between all answers.

The bar chart presented in Figure 5.3 presents for each participant the average error rates they had for all questions they had answered to. There were some response differences between the modalities, but they were low. By summing up the answers of all the participants, there were in total 197 response pairs, as 10 participants answered 18 questions, while the pilot participant answered 17 questions due to the bug mentioned in subsection 5.1.1.

The error rate for all responses but one was of 1 point, while there for a single question was an error rate of 2 points. No higher error rates were registered between the answers from the data. The average error rate between the modalities were of 5,5 points, while the average error rate between all the responses were of 0,29 points.

Of the 197 questions that were answered, 60 of the response pairs had an error rate and the average error rate of those pairs was of 1,01 points. The rest of the 137 response pairs consisted of the same answers. This shows that the participants answered the same in both modalities for a majority of the questions. To further investigate if differences between the response pairs were of statistical significance, the Chi-Square test was used to investigate the relationship of the data. The result from the Chi-square test, X^2 (1) = 73,03, p < 2,2e-16, supports that there is a trend that the participants answers the same in both modalities.

5.3 Analysis of Conversation Logs

This section will present patterns which have been detected from analysing the conversation logs from the participants. The conversation logs have been analysed to find response patterns, length of responses, and the number of participants who received fallback messages.

Of the participants, three of them had previous knowledge of the ASRS test. The transcripts reveal that five of 11 responded with a sentence or more to the questions, while the rest of participants had responded by using short keyword-based answers. Two of the participants are in group A, while three are in group B.

The types of response ROB received can be categorized into three categories. Single phrase responses, a sentence with the frequency-based phrase, and a sentence that adds extra substance to the response. The single phrase responses constitute the responses with a response alternative from the Likert-scale, or one of its synonyms. In the responses from those who wrote in sentences are there some responses that may add extra substance to the test result, but there are also many of them that does not provide additional information. For instance, does some participants mention that they may forget appointments or obligations if they are said verbally, but that they find them easier to remember if they are given in a more formal format. There are though more responses where the users in practice only uses a single frequency-based word, and then wraps it into a sentence without adding more substance to the answer. Here for instance by answering to a question in a manner like “I rarely do that”. To illustrate what types of answers the participants had for the questions, are here some examples for each category that was established:

How often are you distracted by activity or noise around you?

Example 1: Not very often.

Example 2: I sometimes get distracted by noise or activity around me.

Example 3: Depends on the noise, but sometimes. Conversations are often focus killers.

Regarding fallback messages, the transcript reveals that six of the 11 participants experienced one or more fallback messages in the conversation with ROB. The fallback messages in most cases was quickly understood and made the user get back on track. A participant who attempted to respond to a question with a sentence received a fallback two times, this participant proceeded the conversation by using a frequency-based phrase, as the other information had been noted by ROB. The thread is presented below:

ROB: How often do you interrupt others when they are busy?

Participant: only when it is critical that I need to speak to them

ROB: I'm sorry, I didn't get that. Please answer the question in a degree of how often you experience the given symptom.

ROB: How often do you interrupt others when they are busy?

Participant: only when needed

ROB: I'm sorry, I didn't get that. Please answer the question in a degree of how often you experience the given symptom.

ROB: How often do you interrupt others when they are busy?

Participant: Sometimes

The thread above does also illustrate that the participant interprets the question broader in the conversational interface than in the paper-based modality, where the Likert-scale sets the boundaries of what one could respond to a question.

Figure 5.3 presented the average error rates to the response pairs for all questions. Some questions had higher error rates tied to them than others. A potential pattern was detected, where potentially how a question is written can effect how a user responds to the question.

Question 6 had the highest error rate:

"How often do you feel overly active and compelled to do things, like you were driven by a motor?"

While question 12 had the lowest error rate:

"How often do you leave your seat in meetings, or other situations in which you are excepted to remain seated?"

By looking at the questions one could argue that question 12 is a more specific questions than question 6, which is a broader question that could be more open for interpretation. This is merely an observation, further investigation is necessary in order to establish if there in fact is a correlation to how a question is written and how a user respond to it.

5.4 Analysis of Interviews

How does the participants experience the conversational interface?

To answer this question, this section will present the views the participant had of ROB, based on what was said in the interviews. The interviews were short and had an emphasis on getting insight into the participants' perception on answering the questions of the ASRS test, the aspects of doing it in a conversational interface, the visual and structural design of ROB, and lastly what modality they preferred.

5.4.1 About the ASRS-test

As the participant had completed the ASRS test in two different modalities the first topic of the interviews revolved around how it was to respond to the type of questions that were presented in the test. How the participants perceived the questions and if there were any differences between the modalities they wanted to highlight.

The participants thought on a general note that is was fine to answer the questions from the ASRS test. Three participants said that they thought it was interesting to respond to the questions, due to the nature of the questions. The questions made them think and reflect on their own behaviour in a way they were not used to.

The participants were asked if they had any thoughts about any differences when responding to the questions in the two modalities. Several of the participants thought that by writing an answer to a question by getting to use own words, this made it possible to give reflections that were not possible to give in the schema modality. Participant #3 said *"It did in a way add an extra dimension by writing and having it presented in a known chat interface that one is used to. It felt a bit more personal in a way".* Some participants did also mention that they thought they answered differently to the questions from modality to modality.

While all participants thought that it was fine to respond to questions in the ASRS test, two participants raised some critical remarks around the formulation of the questions. Participant #3 found some of the questions to be a bit cumbersome formulated, the participant had to read some of the questions repeatedly to get the essence. #3 added that his mental state of the day could have been the cause to this issue. Further, participant #9 suggested that the questions could have been more customized and contextualised for a conversational interface.

The schema presented all the questions at once. According to some participants, it was easier to read the questions in the conversational interface, due to there being one question presented at the time. On the other hand, it was more obvious for the participants what to respond in the schema due to having all the response alternatives available in a Likert-scale ranging from "never" to "very often". It also felt faster to accomplish the test in the schema.

5.4.2 Responding to the Questions

Further the participants were asked if they found it easy or challenging to formulate a response in the conversational interface, and if they had received any fallback messages.

All the participants generally found it easy to respond to the questions. Some did mention that there were a few seconds where one had to think and find a right word or phrasing. As mentioned five of the 11 participants did generally write in sentences. In group A, where most of the participant had no prerequisite knowledge of the ASRS test, did the participants find it easy to find a proper response. Participant #3 did not respond with full sentences. #3 thought by the way the questions were formulated, by starting with "how often", it encouraged #3 to write short and concise responses by the use of keywords as "not often" and "sometimes". #3 argued *".. it was purely intuitive for me to respond with the two".* This did give insight of why many of participants did respond with short answers.

Participants #2 and #5, who wrote with full sentences found it easy to respond to the questions. They did both also mention that after answering a few questions that they began to see the logic behind the chatbot, that in which ROB looked for certain keywords or phrases. #2 and #5 both said this influenced how they responded in the later questions by shortening their answers.

Among the participants in group B and also participant #1, who had completed the ASRS-test in the schema before the chatbot. They all said it was easy to respond to the questions, but that it also could have been because of having previous knowledge of the Likert-scale

used in the schema. Three of the participants did mention that their responses were influenced by the responses from the schema.

The length of the responses was a topic which were brought up by two of the participants. Participant #10 was a bit uncertain about this aspect and suggested that ROB could send an instructive message giving the user a preferred response length to a question.

Regarding fallback messages, the participants who received them found them to be understandable. They understood that they had forgot to mention the time aspect in their response. Not much more was said about the fallback messages, besides a comment from participant #9 where the participant said he interpreted the message in the fallback as ROB wanted to receive a frequency-based response.

5.4.3 Responding Openly to Questions

The participants further were asked about how they felt about answering and having the opportunity to the questions in an open manner, and if it led to more or less reflection around the questions. The participants liked having the opportunity to write more broadly around the questions. A common response was that by having the opportunity to respond openly one can easily add contextualised information to a response, which is harder to in a schema. Participant #2 argued that this was useful when responding to question where the context may adjust the response. #2 exemplified this by mentioning the question regarding the case of disturbing other people in a work setting, *"this can be necessary sometimes, so I did answer that I do this sometimes. I wrote the reason for this in the prototype. I did not get to explain this in the other test".* Two participants mentioned that they shortened their answers when they got a sense of what ROB looked for in a response, as in patterns and keywords.

The conversational interface did according to most participant lead to more reflection around the questions. Five participants wrote in sentences, but according to some of the participants due to necessity of having to write the answer instead of setting a cross, did it lead to more inner reflection before writing the response.

There were also some critical remarks around this. When asked about how it was to be able to answer openly did participant #9 say *"Of course it was more than one of five responses, but I did use the same vocabulary as the paper version had".* #9 did also point out that when asked about if the test led to more reflection that it led to some reflection, but the questions could have been changed and customized more for a conversational interface. #9

suggested that ROB could have been more contextual around the responses, because ROB in its current form only gave feedback if something was wrong or by asking a new question. *"So I felt it was a bit like, I did just respond to a schema in a chat format".* When asked about if more personality could have improved the experience #9 responded that *"it could be a possibility".*

5.4.4 Feedback on the Design

A central topic of the interviews was the design of ROB. The participants came with positive and critical remarks around the visual and conversational design of the prototype.

Beginning with the visual design, all the users thought the design was easy to understand and found it similar to other chat interfaces they had used. But there were also some critical comments to the design. Six of the participants brought up one design issue that they found problematic. The input field where the users wrote their responses was a bit hard to detect at the beginning, participants thought it could have been more visible. It was also suggested that the interface could have had used more of the screen, due to the large whitespace. Lastly, it was suggested that, ROB could adapt some similar visual elements that are known from other popular chat clients, such as a "writing in progress" animation as seen in for instance Facebook Messenger.

Furthermore, it was suggested that ROB could have some additional functionality alongside the screening. When the screening test ended the participants received a score from ROB. If the person though got a positive result, the user did not get any further information. Two participants thought it would be an improvement if ROB could supply them with relevant information about ADHD and how one could get in touch with a domain expert who could help for further evaluation.

A topic which were brought up by several of the participants was the design of the conversation. When a participant successfully responded to a question then ROB asked a new question. There was a wish for a better form of feedback to the address that ROB had received the response to a question. Participant #2 addressed this topic and said *"I felt like, if the point was that you were supposed to be talking with a robot or a, it did then feel like a very cold thing .."*. Suggested solutions to improve the interaction was to for instance that ROB could send a form of visual cue to the user when a response has successfully has been given and small talk between the questions. ROB could for instance thank for a response or say that he understood the response he had received.

To give ROB a more personal and empathic touch it was suggested by some of the participants that ROB could have a visual embodiment either as human or a robot. The figure of ROB could be an animated figure on the sidebar, or a simple image besides the messages sent by ROB. Participant #10 thought this could give ROB more human-like traits and could help to the calm the user if a user was nervous about responding to the questions because of the chance of possibly getting a positive result from the test.

5.4.5 The Participants' Preference

The last question addressed the preference of modalities. The participants were prompted to reflect and make arguments for their preferred interaction modality. Of the 11 participants did 10 prefer the conversational interface, while the last participant did not have any specific preference. The participant who did not have any preference thought that both modalities were easy to understand and thought that the conversational interface was a bit livelier experience in comparison to the schema but found the schema faster to complete. The participant did not have any strong preference for either modality.

The other 10 participants who preferred the conversational interface argued for their case by referring to previous arguments. Common arguments were that the conversational interface opened for giving more information in a response by making it possible to use own words and sentences, without being strictly bound to the words provided from the questionnaire. Participants thought it was positive that one could reflect around a scenario to give more depth to a response where it was necessary. Participant #5 said the result from chatbot felt more "*right"* and more representative than the result from the schema because of arguments mentioned.

Participant #9 was, as mentioned, critical about how open the test was in the conversational interface, but the participant thought that it was positive that one could give reflections in a response. The participant exemplified this by bringing up an example regarding the questions about taking turns. *"When I play a game or something, I can finely wait for my turn, but if it is a queue .. then it is different."*.

While the conversational interface was preferred, were there also some critical concerns about the modality, for instance the visual and conversational design. Another topic that was brought up in the end was validity of the test result from the chatbot. Participant #2 preferred the conversational interface but added that he would be a bit sceptic about the result from ROB, if that result was one of the main factors leading to a diagnosis. Though

did this get less of an issue when it was explained that that the result from the ASRS test were of a guiding character and not a diagnosing one.

5.5 Summary of Chapter

This chapter has described the evaluation of ROB by presenting the experiment that was conducted and the data that was retrieved from the data, including the result differences between the modalities and the user's experience of ROB.

Chapter 6

Discussion

The thesis has thus far described the design, development and evaluation of ROB, this chapter will present a discussion of the work up against the research question:

How can we design a conversational interface for the ASRS test?

To respond to this research question, a prototype for a chatbot, ROB, was designed and implemented as it has been documented in Chapter 4. The development of the prototype has been inspired by literature and other chatbots that were presented in Chapter 2. The prototype was developed and evaluated by using the methods that were described in Chapter 3. After the development of the prototype was finished, a comparative experiment was conducted. There were two objectives for conducting the experiment. To compare the results from the modalities and to get insight into how the participants experienced ROB in comparison to the paper-based modality.

This chapter has its focus on discussing the methods that has been used in the development of the prototype, the results from the user experiment, and to discuss the prototype itself.

6.1 Discussion of Research Methods

The objective of this study has been to design a proof-of-concept prototype to explore how a screening test could be designed to a conversational interface. Research through design has been used an overarching framework to guide the process of designing the prototype, and to gain knowledge from the process and the prototype. Following the framework, a literature review was conducted to get an overview of literature and relevant work, to justify the relevance of this study. Furthermore, as the development process began has been documented (see Chapter 4) in detail to give a justification for the decisions that has been made in the research.

For the evaluation of ROB, it was determined to conduct a controlled comparative experiment, since the prototype presents the Adult ADHD Self-report Scale (ASRS) in a new modality. To study the potential of the conversational interface, 11 participants were

given the task of completing the ASRS test in a conversational interface and in a paper-based modality. There were two motivational factors for conducting the experiment. The first factor was to find out if participants would answer differently between the modalities. The answers to each question and the results from the ASRS test has been compared for this purpose. The ASRS test has a good validity (Adler et al., 2006; Silverstein et al., 2018), if the responses and scores then would have been of a significant difference, it could set the validity of the result to question. Secondly, it was of interest to find out how the participants experienced completing the ASRS test in a conversational interface.

To get insight around the participants experience of ROB, a semi-structured interview with each participant was conducted as they had finished the tests. These interviews gave the participants the opportunity to talk and reflect around the topics of the interview. The interviews collected valuable information around the participants experience of using ROB.

6.1.1 Research Limitations

It is not unusual that there may be limitations in research, and this research is not an exception. There a few limitations of the study that are necessary to bring up for discussion.

First, the number of participants is a limitation for the study. 11 participants were recruited for the experiment, they provided valuable information for the study. The data from the study would though have had a greater statistical validity if the population of the test group was larger and more distributed. Having more participants though would have required more time and scheduling.

ROB as a chatbot is built to screen a user for ADHD symptoms, but in the user-experiments did none of the participants have ADHD. It is not given that a user with ADHD would have had other types of comments than the ones from the recruited participants. Though, from a research perspective it would have been interesting to get more insight on how participants with the diagnosis perceived the prototype and if the result differences would have been any different.

As mentioned in chapter 4, the English version of the ASRS test was used as the fundament for the development of ROB. None of the participants that were recruited for the experiment though were native English speakers. The participants had sufficient knowledge of the English language to complete the tests, but it could be more challenging to formulate answers for a test like the ASRS in a second language.

The last limitation of the study which is relevant to bring up is tied to the learning effect in the experiment. The participants who completed the ASRS test in the schema before the chatbot, they either had a negative error rate or not a difference at all between the results (see subsection 5.2). While in the group of participants who completed ASRS in the chatbot, they had a higher average error rate than the other group. In the experiment described in the study, the participants were divided into two sub-groups ordered after the modality they were exposed to first. It was designed in this way in order to study the learning effect. Additional measures could have been done to study the learning effect. An alternative way of doing it, it could be by having a participant complete a test in a modality one day and the other modality later. Furthermore, the modality of the test does not affect the symptoms which a participant has or not. It is common to complete a test like the ASRS test multiple times over longer time periods, since the way one experiences symptoms may be relative and may change over time. Therefore, it could also have been interesting to evaluate this aspect to further detail. Unfortunately, to do this it would have required more time and scheduling.

6.2 Discussion of the Research Results

This section will discuss the results of the research by discussing the design process and the user experiment up against the research objectives of the study.

Following a research through design approach, a literature review was conducted to get an overview of relevant literature and work (see subsection 2.4. There exists chatbots for symptom-checking, but none of chatbots that were reviewed were built by using a symptom-check test such as the ASRS. Three chatbots for symptom-checking were presented, but these chatbots had other approaches and other priorities in its implementations. The symptom-checking was more general as it looked for a problem based on the basic symptoms that was described to the chatbot either by text interpreted AI or answering yes/no question to a sequence of questions. Else, little research had been conducted on how to design a screening test to a conversational interface. The proof-of-concept prototype of this study had its focus on designing the ASRS into a conversational interface.

In the design process of ROB, there was a vision for how it could be designed. Since the ASRS was supposed to be designed for a conversational interface, it was of interest to try an adjust the test to a conversational format where users could write their answers, instead

of pushing buttons. Using buttons for the design was something which was under consideration early in the design process. Buttons could have made the interaction easier and quicker for a user. It was though decided not to use buttons as a primary input, since it would have made ROB to similar to a traditional questionnaire. Text was therefore used as the primary input for the screening dialog. In appendix C, there are two wireframe mock-ups of a suggested design with buttons, and a screenshot of an early prototype of ROB implemented with buttons.

A high-fidelity prototype, ROB, was designed for presenting the ASRS in a conversational interface. ROB was designed so users could answer more open to questions than compared to the traditional paper-based modality. By designing and developing a functional high-fidelity prototype of ROB and testing it with participants, there has been gained valuable knowledge that has been crucial in the matter of answering the two sub questions that were defined for the research.

From the experiment it was reported that there were individual result differences between the modalities. There also were some differences between how the participants responded to each question between the modalities, as Figure 4.1 shows. This could be the result of a learning effect, where the participants may have had changed their response when they answered a question for a second time. The changes were through small as the most common error rate was of 1 point, something which did not drastically change the symptoms for the participant. The symptoms a person has can't change much in the short time-span the experiment was conducted in. Frequency-based words on the other hand as the ones used in ASRS could be interpreted in different ways, something which also can make a person adjust an answer when taking a test like the ASRS a second time.

The results from the first part of the ASRS test was analysed to investigate the result differences between the modalities. Further, the answers to all questions were investigated by using the Chi-square test (see subsection 5.2). The test was used to investigate the relationship between the responses to each of the questions. The result from the Chi-square test supported the notion there was a trend where the participants answered the same to the questions in both modalities. There were 60 response pairs where the responses were different from each other, but the error rates for all these pairs were small. The mean error rate of those response pairs was of 1,01 points. These results are positive as they show that the results provided by the conversational interface does not have a large error rate from

paper-based test. This may also support the belief that the result from ROB could be valid, but as mentioned should the result be interpreted with caution due to the size of the participant group.

From analysing the quantitative data, it was also revealed a pattern which suggests that the way a question is written it could influence a response in the conversational interface. The pattern implied that there were larger error rates to questions that were more broad and open for interpretation, than the questions that were more specific. It could imply that participants interpret these questions differently when they are not bound to using the answers from the Likert-scale. Unfortunately, it was not investigated further in this study if there is a correlation between those factors.

The ASRS test is as mentioned the first step towards a further evaluation, the result from the test gives an indication of whether a patient shows symptoms for ADHD or not. The patient either does complete the test as a questionnaire or as in a conversation with a domain expert. As it is common to use the ASRS as an interview guide for a domain expert, this was a motivational factor for exploring how a chatbot can possibly simulate a conversation like this with a patient if there are concerns tied to having symptoms of ADHD. 10 of 11 participants preferred the chatbot to the paper-based schema, regardless of the limitations that ROB had. Additionally, the participants used a bit more time on completing the tests with ROB, but the responses from the conversational interface could collect more details around a symptom in comparison to the responses from the schema. The aspects of the prototype itself is discussed in the next section.

Judging from the results could one argue that there is an indication that a chatbot can be a useful screening utility in the mental health domain. As most users answered the same in both modalities, it could imply that the method of using synonyms for the dialog design has been successful, regardless of the errors that were reported in subsection 5.2. The results would have been acceptable if ROB had the word that was missing from his thesaurus, as the average error rate was low. Additionally, a majority of the participants favoured the conversational interface. The results argue for the case of using chatbots as a utility for screening. A chatbot could easily be deployed to a webpage or to a messaging platform such as Facebook Messenger. This could widen its reach and make it more accessible for people who may experience symptoms of ADHD but are uncertain if they want to contact a domain expert for their eventual problem. Furthermore, if the chatbot is connected to

system connected to domain experts as ADA or Babylon (see subsection 2.4.3, the results could potentially be used further by the experts in a further evaluation.

6.3 Discussion of the Prototype

This section will discuss the design of the prototype and present design implications which has been elicited from the discussion. The section has a focus on the conversational structure and the role synonyms had in the design of ROB.

6.3.1 The Design of the Prototype

The capabilities of the prototype in its current form is limited due to the fact it has been designed to explore how the ASRS test could be designed for a conversational interface. Because of the limited scope for this study did it leave out for instance having ROB respond to questions about the ADHD diagnosis or other practical questions. For potential future developments ROB could be a part of an another chatbot with more functionality, such as an assistant chatbot which can provide the mentioned functionality and more.

The results from the comparative experiment indicates that the participants had a good perception about completing a screening test such as the ASRS test in a conversational interface. A goal that were guiding the design of the prototype was to make a minimalistic design with few distractions, following one of the guidelines by Sonne, Marshall et al. (2016) (see subsection 2.4.1). In contrast to the paper-based modality, ROB asks a user one question at the time. From the interviews (see subsection 5.4.1), it was revealed that participants experienced this way of presenting the questions as tidier, as ROB presented one question at the time and the schema presented all questions at once.

The elements of the visual design that were criticised in the interviews (see subsection 5.4.4) are simple to adjust, but remarks that are worth discussing more in detail are the ones pointed towards the dialog structure of ROB. As ROB is a chatbot, the perception of the dialog is important. A central critical remark that repeated itself among the participant was the nature of the conversation with ROB. The conversation was found to be a bit cold and less empathetic than it should have been. It was question upon question, without any comments in between the question, in which reduced feeling of it being a conversation. By having such a functional structure did some of the participants mention uncertainty around if ROB had received their response. Small comments and visual effects were suggested to handle this problem. This criticism is valid, as goal while making a chatbot is often to make the chatbot simulate a conversation one could have with a human (McTear et al., 2016a).

If the conversation with ROB was to simulate a conversation with a domain expert, it is indeed not sufficient as it is in its current form.

As it was reported in Chapter 5, some of the participants who answered with sentences, they shortened their answers when they got a sense of what ROB was looking for in a response. To finish the test with ROB, it is sufficient to answer with one of the alternatives from the Likert-scale. From the results reported in Chapter 5, it was implied that it was limited how much the current conversational design engaged the participants to write more informative answers to ROB. When designing a test like the ASRS for a conversational interface, a new modality, it opens for new ways to gather data regarding symptoms as users can write and add more information to a response. Five of 11 of the participants supplied their answers with more information, while the other participants answered with short answers. One possible explanation for this is the design of ROB, where he asks questions which starts with "how often". For a potential future improvement, an ideal solution would have been to make the screening test more customized for a conversational modality. If there were different questions, or if they were asked differently than "how often", it could perhaps have made the participants also write more in their answers. On the other hand, it could also have a trade-off. The structure and the design of the ASRS test is what gives the test its validity. If the questions of the test were changed, it could possibly reduce the tests attributes which makes it a useful utility for screening. The structure of the ASRS test was not changed in the design of Rob, due to the wish of not inflicting the validity of the test. To make users write more, another possible solution to enhance the conversation could be to make ROB ask contextual follow-up questions to the questions from the ASRS test. In the interviews from the evaluation it was also mentioned that ROB could have had come with small comments between the questions, comments which could have prompted a user to write additional information. Writing more in a response may not always be necessary but having the opportunity could one argue is one of the advantages a conversational interface may offer in comparison to a questionnaire.

6.3.2 Using Synonyms for the Design

The proof-of-concept prototype developed for this research was designed to look for keywords and its synonyms from the input it received. From the results of the comparative experiment, it was proven to be an intuitive and sufficient method to design a conversational interface for the ASRS test. Using this synonym-based approach, it enabled users to write open answers if they used one of the words from the Likert-scale or one of

its synonyms. As a question is asking for a frequency-based response, it was according to the participants simple to find the right words that was necessary to proceed in the test. Though as the words in the vocabulary was added manually, this also was a weakness for this method of designing a chatbot. As it was mentioned in subsection 5.3, there were a few participants who had written “not very often’ a few times as a response for the questions. ROB did not have this phrase in his vocabulary, which in return led to the response being interpreted as “very often”. This flaw gave participant #1 a score of 19, a score 6 points higher than what is was supposed to be because of this flaw (see Table 1). The flaw was simple to fix, though it may be a problem if simply a negation of a word or its synonym could lead to a false result.

On the other hand, if ROB had the missing word in his vocabulary, the results would have been fine since the error rates were few and small. To improve the structural design of the prototype, it could have been conducted a more thorough job with manually adding synonyms to ROB’s thesaurus in order to prevent misinterpretation of negations of words. When using a synonym-based approach like the one presented in this study, one should also consider what to do if the input has multiple opposing entities. For the design of ROB, it was considered to have him re-ask questions if this happened. Unfortunately, it was not implemented for this prototype due to uncertainty of implementation (see subsection 4.5.4).

6.3.3 Design Implications

Below are three design implications which has been elicited from the chapter’s discussion for future research regarding how to design a conversational interface for screening.

1. Consider having small comments in between questions, as the conversational screening dialog may be perceived as cold and less empathic without it.
2. Consider outlining questions that are not frequency-based if there is a wish for longer answers.
3. Precautions should be made if using a synonym-based method for design. If a chatbot lacks words or negations of words in its thesaurus, it could lead to misinterpreting the intention of the user’s input (see subsection 6.3.2).

6.4 Chapter Summary

This chapter presented a discussion of the research by discussing the methods, the results from the experiment, and the high-fidelity prototype that has been implemented. In the end

were some design implications suggested for further research on how to design conversational interfaces for screening.

Chapter 7

Conclusion

The research presented in this thesis has studied how a screening test could be designed to a conversational interface. The motivation for the research was the objective to explore how one could make use of conversational interfaces for screening in the mental health domain.

Following a research through design approach. a high-fidelity prototype was developed through three development phases. The result was ROB, a screening chatbot which presents the Adult ADHD Self-Report Scale (ASRS) in a conversational interface. In contrast to the regular questionnaire, ROB gives users the option to respond to the questions from the test with more open language, instead of being bound to five response alternatives. This gives the user the option to supply responses in the test with more information around a symptom, information which a user could find it relevant for domain experts to know of.

A comparative experiment was conducted, where 11 participants completed the ASRS with ROB and in the paper-based version of the test. The experiment had two objectives. First, to compare the answers and the results between the modalities. Second, to get insight into how the participants experienced using a chatbot for a screening test. If the results had a larger error rate between the modalities, it could set the validity of the result from ROB to question. From the results, it was reported there was a trend for the participants to answer the same in both modalities. There were a few individual response differences, but the error rates between the modalities were few and small. This supports the notion that the results from ROB could be valid, and that a conversational interface is something which can be used as a utility for screening. Furthermore, the participants used more time to complete the test with ROB, but in return did ROB in some instances receive responses which gave more information to the result. Many participants though responded to the questions by answering with short keyword-based answers, so a test like the ASRS itself is not adjusted to a conversational interface in its current form, if there is a demand for more informative responses from a conversational interface.

In addition, it was reported that 10 out of 11 participants favoured the conversational interface to the schema, as the conversational interface felt more personal and engaging. It also gave the participants the option to supply their answers with more information where they felt it was necessary. Participants did though have critical remarks towards the design of ROB, where a central topic was the design of the dialog, where it can be boiled down to the screening could have been more customised to a conversational interface, by having ROB ask contextual questions and come with small comments between questions.

7.1 Future Work

The functionality ROB provides at this point is limited, since it is a proof-of-concept prototype designed to study how the ASRS could be implemented to a conversational interface. For potential further developments, ROB could either expand his feature set or be implemented into an another chatbot. This could for instance be a conversational assistant tailored to assist adults with ADHD. Furthermore, to reach target users who may experience symptoms of ADHD, ROB could be deployed to a known messaging platform such as Facebook Messenger or Skype. This could widen the reach for the chatbot to make it more visible and accessible for people who may experience symptoms of ADHD.

ROB's design revolved around using synonyms for proceeding the conversation. It worked out for this study. For potential future developments, it could be further developed in order to address the shortcomings of the current implementation. For instance, by making ROB better to detect negations of entities or to detect opposing entities.

For potential future research on conversational interfaces for screening, it would have been interesting to evaluate ROB amongst more users to strengthen the statistical validity of the results. There are also other data points to research that has not been covered properly in this thesis. For instance, what causes different answers to a question in the modalities. It was briefly mentioned that more concrete questions had a lower error rate than compared to questions that were more open to interpretation. Future research could investigate this aspect further.

Bibliography

ADA. (n.d.). Ada - Personal Health Companion App. Retrieved May 2, 2018, from https://ada.com/

ADHD Norge. (2016a). Hva er ADHD? – ADHD Norge. Retrieved March 10, 2017, from http://adhdnorge.no/voksen/hva-er-adhd/

ADHD Norge. (2016b). Medisinering av ADHD – ADHD Norge. Retrieved May 30, 2017, from http://adhdnorge.no/voksen/medisinering-av-adhd/

Adler, L. A., Spencer, T., Faraone, S. V., Kessler, R. C., Howes, M. J., Biederman, J., & Secnik, K. (2006). Validity of Pilot Adult ADHD Self- Report Scale (ASRS) to Rate Adult ADHD Symptoms. *Annals of Clinical Psychiatry*, *18*(3), 145–148. https://doi.org/10.1080/10401230600801077

Atlassian. (n.d.). What is Git: become a pro at Git with this guide | Atlassian Git Tutorial. Retrieved March 10, 2018, from https://www.atlassian.com/git/tutorials/what-is-git

Barkin, J. (2016, August 23). How to avoid UI dead ends when building your chatbot | VentureBeat. *VentureBeat*. Retrieved from https://venturebeat.com/2016/08/23/how-to-avoid-ui-dead-ends-when-building-your-chatbot/

Bayazit, N. (2004). Investigating Design: A Review of Forty Years of Design Research. *Design Issues 20, No*. The MIT Press. https://doi.org/10.2307/1511952

Biederman, J., & Faraone, S. V. (2005). Attention-deficit hyperactivity disorder. *The Lancet*, *366*(9481), 237–248. https://doi.org/10.1016/S0140-6736(05)66915-2

Bødker, S. (2006). When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction changing roles - NordiCHI '06* (pp. 1–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/1182475.1182476

Bødker, S. (2015). Third-wave HCI, 10 years later---participation and sharing. *Interactions*, *22*(5), 24–31. https://doi.org/10.1145/2804405

Braut, G. S. (2018). screening – Store medisinske leksikon. Retrieved March 15, 2018, from https://sml.snl.no/screening

Brown, T. E. (2008). ADD/ADHD and Impaired Executive Function in Clinical Practice. *Current Psychiatry Reports Current Medicine Group LLC ISSN*, *10*, 407–411. Retrieved from https://pdfs.semanticscholar.org/37e3/6d6c17116debfa61413b52fbcee5615f51f0.pdf

Bu, E. T. H., Skutle, A., Dahl, T., Løvaas, E., & van de Glind, G. (2012). Validering av ADHD-screeninginstrumentet ASRS-v1 . 1 for pasienter i behandling for rusmiddelavhengighet. *Tidsskrift for Norsk Psykologforening*, *49*, 1067–1073.

Churchman, C. . W. (1967). Wicked Problems. *Management Science*, *14*(4), B-141-B-146. https://doi.org/10.1287/mnsc.14.4.B141

Cross, N. (1982). Designerly ways of knowing. *Design Studies DESIGN STUDIES Vol*, *3*(3), 221–227. https://doi.org/10.1016/0142-694X(82)90040-0

Cross, N. (2001). Designerly ways of knowing: design discipline versus design science. *Design Issues*, *17*(3), 49–55. https://doi.org/10.1162/074793601750357196

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., … Morency, L.-P. (2014). SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (pp. 1061–1068). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. Retrieved from http://dl.acm.org/citation.cfm?id=2617388.2617415

Fischer, M., & Lam, M. (2016). From Books to Bots: Using Medical Literature to Create a Chat Bot. https://doi.org/10.1145/2933566.2933573

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), e19. https://doi.org/10.2196/mental.7785

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, *24*(4), 38–42. https://doi.org/10.1145/3085558

Følstad, A., Brandtzaeg, P. B., Feltwell, T., Law, E. L.-C., Tscheligi, M., & Luger, E. A. (2018). SIG. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (pp. 1–4). New York, New York, USA: ACM Press. https://doi.org/10.1145/3170427.3185372

Gregory, S. A. (1966). Design Science. In S. A. Gregory (Ed.), *The Design Method* (pp. 323–330). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-6331-4_35

Harrison, S., Tatar, D., & Sengers, P. (2007). The Three Paradigms of HCI. *CHI*. Retrieved from http://people.cs.vt.edu/srh/Downloads/TheThreeParadigmsofHCI.pdf

Helsedirektoratet. (2014). ADHD/hyperkinetisk forstyrrelse – nasjonal faglig retningslinje for utredning, behandling og oppfølging\rRett diagnose – individuell behandling. *Nasjonale Faglige Retningslinjer*. Retrieved from https://helsedirektoratet.no/Retningslinjer/ADHD.pdf

Hevrøy, H. O. (2016, October 4). Har levd lenge med ADHD uten å vite det – Livsstil. *NRK*. Retrieved from https://www.nrk.no/livsstil/xl/har-levd-lenge-med-adhd-uten-a-vite-det-1.12922276

IBM. (2018). Defining entities. Retrieved May 13, 2018, from https://console.bluemix.net/docs/services/conversation/entities.html#defining-entities

Intromat. (2016). Cognitive training in ADHD – Intromat. Retrieved March 10, 2017, from http://intromat.no/cases/cognitive-training-in-adhd/

INTROMAT. (2017). About INTROMAT – Intromat. Retrieved May 24, 2018, from http://intromat.no/about/

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., … Walters, E. E. (2005). The World Health Organization Adult ADHD Self-Report Scale (ASRS): a short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15841682

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction* (Second Edition). Morgan Kaufmann. Retrieved from https://www.sciencedirect.com/science/book/9780128053904

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA" In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 5286–5297). New York, New York, USA: ACM Press. https://doi.org/10.1145/2858036.2858288

MacKenzie, I. S. (2013). *Human-Computer Interaction: An Empirical Research Perspective* (1st ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

McTear, M., Callejas, Z., & Griol, D. (2016a). Conversational Interfaces: Past and Present. In *The Conversational Interface* (pp. 51–72). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32967-3_4

McTear, M., Callejas, Z., & Griol, D. (2016b). The Dawn of the Conversational Interface. In *The Conversational Interface* (pp. 11–24). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32967-3_2

Mozilla. (n.d.). Working with JSON - Learn web development | MDN. Retrieved April 9, 2018, from https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Objects/JSON

NinjaMock. (n.d.). NinjaMock. Retrieved March 10, 2018, from https://ninjamock.com/

Nordea News. (2017). Nova er bankens mest effektive medarbeider - Nordea News / Nordea News. Retrieved May 2, 2018, from https://nordeanews.no/2017/09/hun-er-bankens-mest-effektive-medarbeider/

O'Hear, S. (2017a, April 19). Ada is an AI-powered doctor app and telemedicine service | TechCrunch. *Techcruch*. Retrieved from https://techcrunch.com/2017/04/19/ada-health/

O'Hear, S. (2017b, April 25). Babylon Health raises further $60M to continue building out AI doctor app | TechCrunch. *Techcrunch*. Retrieved from https://techcrunch.com/2017/04/25/babylon-health-raises-further-60m-to-continue-building-out-ai-doctor-app/

Oulasvirta, A., & Hornbæk, K. (2016). HCI Research as Problem-Solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 4956–4967). New York, New York, USA: ACM Press. https://doi.org/10.1145/2858036.2858283

Perez, S. (2017, March 3). Domino's now lets you order from its full menu via Messenger – no setup or account required | TechCrunch. *Techcrunch*. Retrieved from https://techcrunch.com/2017/02/03/dominos-now-lets-you-order-from-its-full-menu-via-messenger-no-setup-or-account-required/

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. https://doi.org/10.1145/3173574.3174214

Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied Conversational Agents in Clinical Psychology: A Scoping Review. *Journal of Medical Internet Research, 19*(5), e151. https://doi.org/10.2196/jmir.6553

Rogers Yvonne, Sharp Helen, P. J. (2011). *Interaction design: beyond human-computer interaction* (3rd ed.). Wiley.

Silverstein, M. J., Faraone, S. V., Alperin, S., Leon, T. L., Biederman, J., Spencer, T. J., & Adler, L. A. (2018). Validation of the Expanded Versions of the Adult ADHD Self-Report Scale v1.1 Symptom Checklist and the Adult ADHD Investigator Symptom Rating Scale. *Journal of Attention Disorders*, 108705471875619. https://doi.org/10.1177/1087054718756198

Sinfield, J. (2018). Understanding ADHD in Adults. Retrieved March 27, 2018, from https://www.verywellmind.com/adult-adhd-4157275

Sonne, T., Marshall, P., Obel, C., Thomsen, H., & Grønbaek, K. (2016). An Assistive Technology Design Framework for ADHD. https://doi.org/10.1145/3010915.3010925

Sonne, T., Müller, J., Marshall, P., Obel, C., & Grønbæk, K. (2016). Changing Family Practices with Assistive Technology. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 152–164. https://doi.org/10.1145/2858036.2858157

Tohidi, M., Buxton, W., Baecker, R., & Sellen, A. (2006). Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06* (p. 1243). New York, New York, USA: ACM Press. https://doi.org/10.1145/1124772.1124960

Trello. (n.d.). Trello. Retrieved March 10, 2018, from https://trello.com/

Vincent, J. (2018, March 20). IBM's Watson Assistant lets any company build Alexa-like voice interfaces - The Verge. *The Verge*. Retrieved from https://www.theverge.com/2018/3/20/17142232/ibm-voice-assistant-watson-b2b-enterprise-interface

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research Through Design as a Method for Interaction Design Research in HCI. Retrieved from http://repository.cmu.edu/hcii

Appendix A

Consent Form

Førespurnad om deltaking i forskingsprosjektet

"Samtalegrensesnitt for symptomsjekktest"

Bakgrunn og føremål
Dette er ein brukartest av chatbotten «Rob». Dette er ein prototype for ein chatbot som har implementert WHO sin symptomsjekktest «Adult Self Report Scale» for ADHD (ASRS). Det er ein symptomsjekktest som blir brukt for å kartlegge symptomar og kalkulere sannsynet for at ein vaksen person kan ha ADHD. Føremålet med testen er å skulle evaluere prototypen og samanlikne resultata ein får etter å ha tatt testen i eit chatformat og i eit skjemaformat. Dette er ein del av ei masteroppgåve i informasjonsvitskap ved UiB, som er under forskingsprosjektet INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology)

Utvalet av testpersonar for studien er studentar i frå institutt for informasjons- og medievitskap.

Kva inneberer deltaking i studien?
Deltakinga er todelt;

1. Først vil du få litt informasjon om prototypen. Du vil deretter ta ASRS testen to gongar. Ein gong i eit chatformat og ein gang i eit tradisjonelt format. Rekkefølga vil variere for enkelte deltakarar. Testansvarleg vil være tilgjengeleg for spørsmål undervegs.
2. Dette vil vidare følgast opp med nokre spørsmål om prototypen du har brukt.

Kva skjer med informasjonen om deg?
Resultatssummane ein får etter å ha gjennomført testane vil bli lagra for å skulle samanlikne dei to ulike interaksjonsformane. Det vil bli utført ei analyse av typen svar som kjem fram i dialogen med Rob, derfor vil samtaleloggen bli lagra. Ingen personopplysningar vil bli samla inn og deltakarar vil bli instruert til å ikkje bruke namn eller andre personidentifiserande opplysningar i dialogen med Rob. Personidentifiserande opplysningar blir fjerna i frå samtaleloggen om dei skulle framkome. Deltakarar vil ikkje kunne bli gjenkjent i studiet.

Prosjektet skal etter planen avsluttast innan 1. juni 2018.

Frivillig deltaking
Det er frivilleg å delta i studien, og du kan når som helst trekke ditt samtykke utan å oppgje nokon grunn.

Dersom du har spørsmål til studien, ta kontakt med Robin Håvik, tlf 41516442 /
Robin.Havik@student.uib.no
Eller vegleiar; Frode Guribye, frode.guribye@uib.no

Samtykke til deltaking i studien

Eg har motteke informasjon om studien, og er villig til å delta

---

(Signert av prosjektdeltakar, dato)

Appendix B

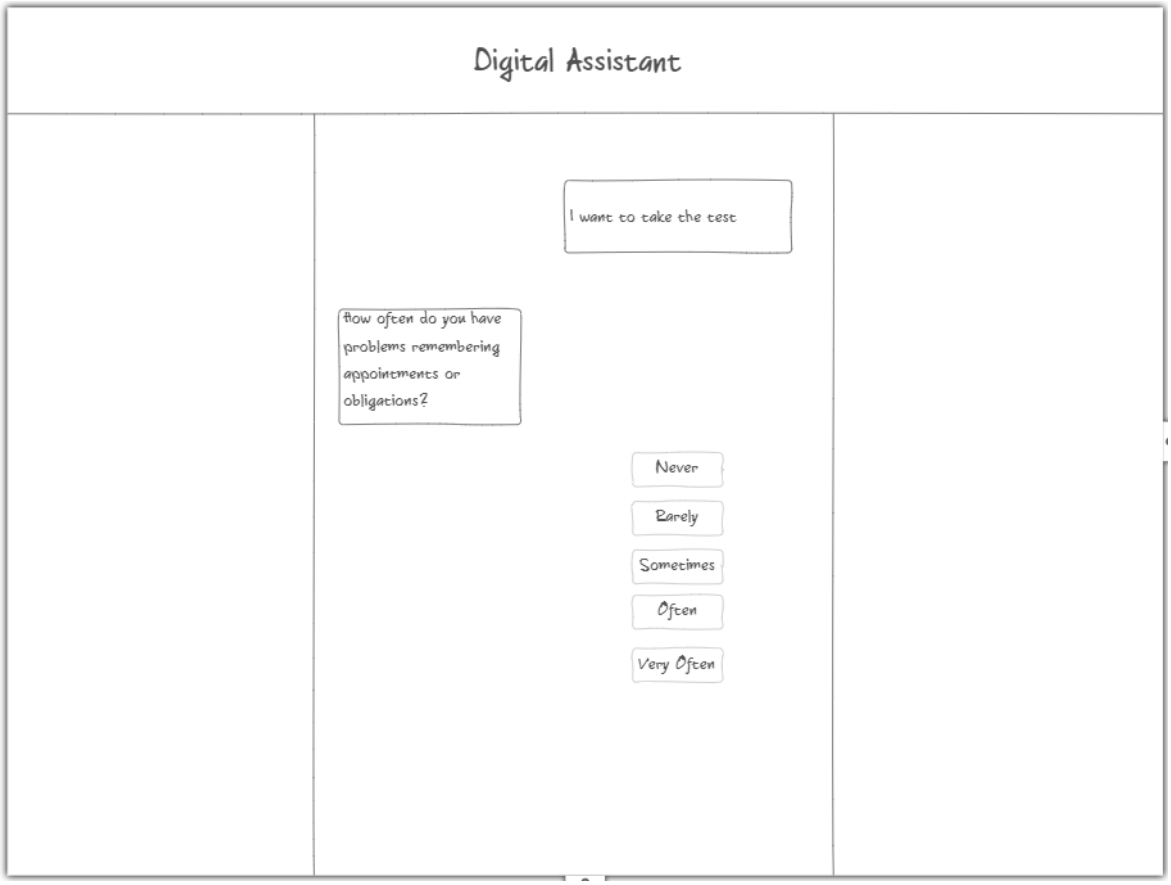
Adult ADHD Self-Report Scale (ASRS)

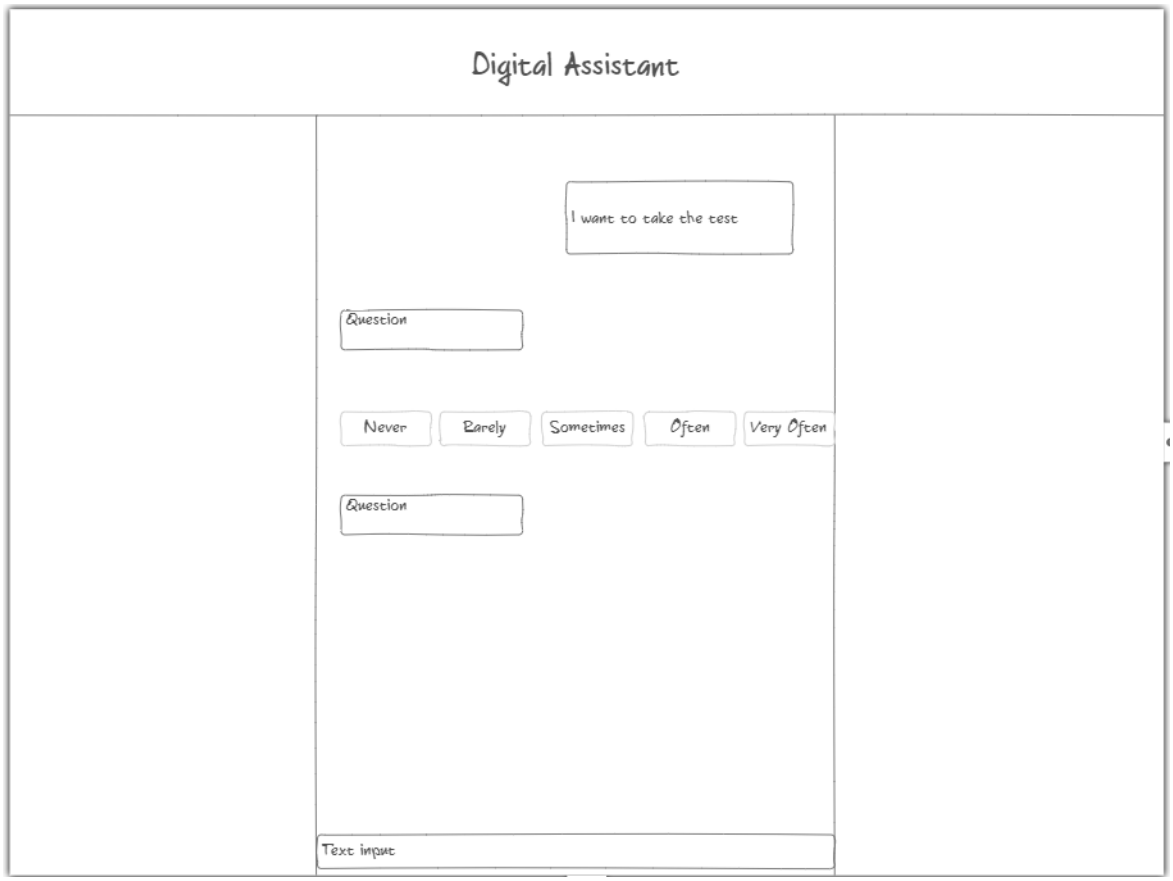
Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist

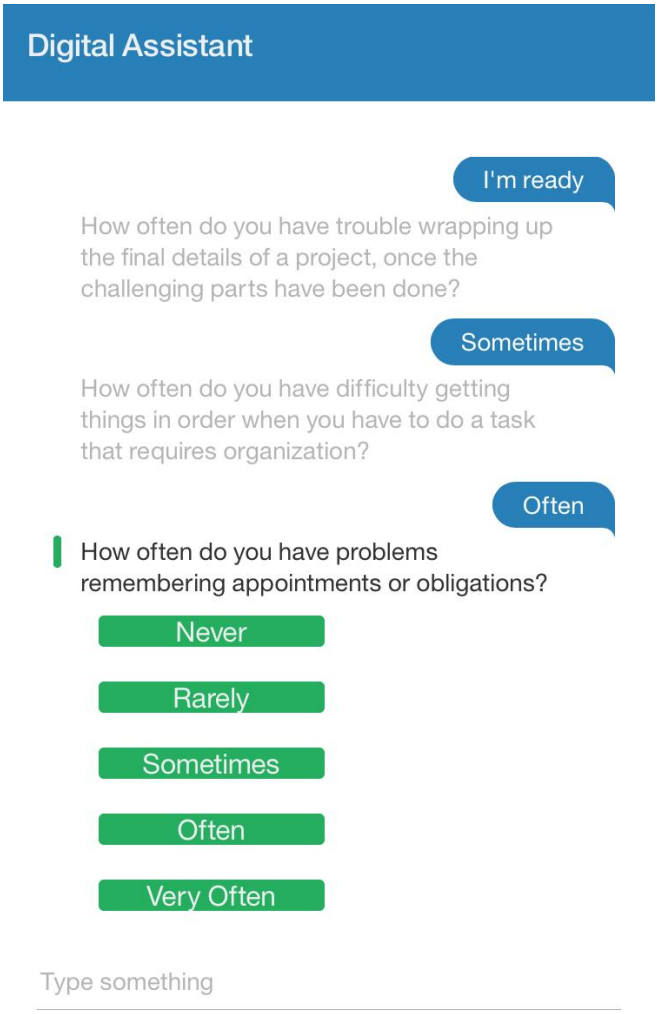
Patient Name		Today's Date					
Please answer the questions below, rating yourself on each of the criteria shown using the scale on the right side of the page. As you answer each question, place an X in the box that best describes how you have felt and conducted yourself over the past 6 months. Please give this completed checklist to your healthcare professional to discuss during today's appointment.			Never	Rarely	Sometimes	Often	Very Often
1. How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?							
2. How often do you have difficulty getting things in order when you have to do a task that requires organization?							
3. How often do you have problems remembering appointments or obligations?							
4. When you have a task that requires a lot of thought, how often do you avoid or delay getting started?							
5. How often do you fidget or squirm with your hands or feet when you have to sit down for a long time?							
6. How often do you feel overly active and compelled to do things, like you were driven by a motor?							
							Part A
7. How often do you make careless mistakes when you have to work on a boring or difficult project?							
8. How often do you have difficulty keeping your attention when you are doing boring or repetitive work?							
9. How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly?							
10. How often do you misplace or have difficulty finding things at home or at work?							
11. How often are you distracted by activity or noise around you?							
12. How often do you leave your seat in meetings or other situations in which you are expected to remain seated?							
13. How often do you feel restless or fidgety?							
14. How often do you have difficulty unwinding and relaxing when you have time to yourself?							
15. How often do you find yourself talking too much when you are in social situations?							
16. When you're in a conversation, how often do you find yourself finishing the sentences of the people you are talking to, before they can finish them themselves?							
17. How often do you have difficulty waiting your turn in situations when turn taking is required?							
18. How often do you interrupt others when they are busy?							
							Part B

Appendix C

Alternative Prototype with Buttons




Digital Assistant
I want to take the test
Question
Never
Rarely
Sometimes
Often
Very Often
Question
Text input


Digital Assistant
I'm ready
How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?
Sometimes
How often do you have difficulty getting things in order when you have to do a task that requires organization?
Often
How often do you have problems remembering appointments or obligations?
Never
Rarely
Sometimes
Often
Very Often
Type something