Eases and Difficulties of Talking to a Virtual Coach for Quitting Smoking and Becoming More Physically Active: A Mixed-Methods Analysis

Arsen Ekinci
Supervisor(s): Nele Albers, Willem-Paul Brinkman
EEMCS, Delft University of Technology, The Netherlands

22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Eases and Difficulties of Talking to a Virtual Coach for Quitting Smoking and Becoming More Physically Active: A Mixed-Methods Analysis

Arsen Ekinci[1]
Supervisor(s): Nele Albers[1] **, Willem-Paul Brinkman**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

Abstract

As the awareness of the risks of smoking tobacco increased, usage declined as people started to attempt to quit smoking more. Currently, a wide variety of mobile health (mHealth) applications focused on smoking cessation exist, however users commonly quit these applications over time. In the case of Sam, a conversational agent that plays the role of a virtual coach that helps people to quit smoking and become more physically active, it is still unclear what eases and difficulties users experience when talking to Sam and what others factors might play a role in this. Using a mixed-methods analysis of data, gathered from participants after using Sam, six themes were identified that could provide an insight into these eases and difficulties. The identified themes were conversations feel unnatural, conversations feel natural, clarity towards the user, comfortable to talk, ease of replying and conversations felt impersonal. Recommendations were given for each team on both what to possibly improve and what to leave unchanged. Though users were positive about talking to Sam in general, there still was room for improvement.

1 Introduction

1.1 Background

Tobacco use, particularly smoking tobacco, is a risk factor for many diseases such as lung cancer [1]. Currently around 80% of lung cancer deaths are caused by smoking and/or inhaling secondhand smoke [2]. In the year 2000, around 33.3% of the global population, aged fifteen years and older, made use of some form of tobacco. By the year 2020 this rate has declined to 22.8%. Even though this number is decreasing, lung cancer is still the leading cause of cancer related deaths worldwide. [2].

With the ever so increasing usage of mobile phones, mobile health (mHealth) could be an effective way to improve the quality of health of the people [3]. Back in 2013, there were over 400 mobile phone applications for smoking cessation for both iOS and Android. Worldwide, these applications were downloaded over twenty million times [4].

Unfortunately, users commonly quit using these mHealth applications over time for a variety of reasons, such as a loss of interest and/or motivation [5]. To this day, it still remains unclear for specific mHealth applications why this occurs due to the lack of existing research. Gaining insight into why users, for example, lose interest and/or motivation amongst other reasons, could allow for identifying possible improvements for the mHealth application(s).

There can be a wide variety of aspects that users can either like or dislike about an mHealth application. An aspect that users tend to experience as positive, when using mHealth applications, is the ease of use of the application. Users enjoy a good flow in the use of the application. On the other side of things, users dislike unreliable, inaccurate and frustrating aspects of mHealth applications [6].

Though related research exists, it often fails to identify specific causes within mHealth applications that are related to the aforementioned positive and negative aspects of using the applications. In the case of frustration, for example, it would be more useful to know the possible causes of this frustration in order to be able to improve the effectiveness of the application in question.

In the case of mHealth applications that mainly consist of chatting with a conversational agent, identifying reasons for finding it easy to difficult to talk to the agent could provide an insight into why users either continue or quit using the application. This is the gap that this research aims to fill.

1.2 Aim

This report will take a look at a specific conversational agent, called Sam, which is a part of the project ”Perfect Fit.” Sam is a virtual coach that aims to help users prepare to quit smoking, as well as help them become more physically active [7]. More specifically, this report will attempt to gain an understanding of possible reasons and factors of why users find it either easy or difficult to talk to the virtual coach in order to answer the following research question: ”What are reasons for finding it difficult/easy to talk to a virtual coach?”.

1.3 Outline

To achieve this, this report first discusses the methodology behind the research. This methodology consists of the qualitative and quantitative analysis performed on the provided data, followed by a literature study. After this, the report discusses method triangulation using these three methods, followed by a section regarding the investigator triangulation performed during the qualitative analysis. Once the methodology has been explained, the found themes and any relevant recommendations to improve the virtual coach are presented. After presenting these results, a section covering the ethical aspects will be included. Finally the report is concluded with the discussion and conclusion of the research. In the conclusion, possible future work and open issues will also be discussed here.

2 Methodology

The overall methodology was split up into three main parts: qualitative analysis by thematic analysis [8], quantitative data analysis and literature study. Furthermore, the thematic analysis was extended by using both investigator triangulation and method triangulation to solidify findings [9]. Investigator triangulation involves the participation of one or more additional researchers to provide multiple observations and insights. Method triangulation involves using multiple methods of analyzing and interpreting data.

For reproducibility and transparency's sake, a GitHub repository was made to store important data [10], other than the originally provided data [11]. This included all interim results such as the final coding, final coding scheme, final themes and any software that was written and used throughout this research. The "readme.md" file in the repository also contains explanations on how to use the written software to reproduce the results.

2.1 The virtual coach

To understand the data, it was important to first understand the application that was used. This application, which was part of the "Perfect Fit" project, allowed users to chat to a virtual coach called Sam [12]. Sam's goal was to help people prepare to quit smoking, as well as help them improve their physical activity. During these conversations, users were able to reply to the conversational agent's messages using both free-text and buttons with provided answers. Sam could for instance ask users to perform some assigned activities, like example writing down their goals, or ask users to rate some statements. There were up to five sessions in total, during which users were able to talk to Sam.

2.2 The data

Before the users were asked to use Sam, they were first required to fill in a pre-screening questionnaire and a pre-questionnaire [11]. The pre-screening questionnaire contained consent checks, questions to see whether the user was part of the target audience of the study and basic data such as age and gender identity. The pre-questionnaire contained questions that provided an insight into the users' characteristics. These characteristics contain, but are not limited to whether users have ever quit smoking for at least 24 hours and personality-related questions.

After the users finished all five conversations with Sam, they were required to fill in a post-questionnaire. This post-questionnaire contained questions that provided an insight into how users experienced their conversations with Sam. For some questions, users were required to rate a statement such as "It was easy to do the assigned activities", on a scale of minus five to five based on how much they agreed with it. For other questions, users were required to give a free-text explanation for their rating.

For this research, pre-screening and pre-questionnaire data was taken into account, the rating users gave regarding how easy and motivating the given exercises were from the post-questionnaire, the rating to the question "How easy or difficult was it talking to the conversational agent Sam?", and the free-text response in which users elaborated on this rating, which were all asked in the post-questionnaire. The four questions about ease and motivation of the assigned activities were also used, but had to averaged as further explained in 2.3. The other five rating questions and their accompanying five elaboration questions from the post-questionnaire were irrelevant and were therefore not used, with the exception of during the familiarization phase of the thematic analysis.

2.3 Pre-processing the data

Before the provided data could be used, it first needed to be pre-processed. During this process, certain values in the pre-questionnaire were reversed and/or averaged into a single value for some multiple item measures. To determine which values should be reversed and which values should be averaged, the provided explanation file was consulted [11]. These averaged values were used during the quantitative analysis, as described in 2.4. These averaged values were the ease and motivation of the assigned activities, which was located in the post-questionnaire. Using the personality questions from the pre-questionnaire, the openness to experiences was calculated and used as described in 2.4 using the calculations by Gosling et al. [13].

Cronbach's Alpha was used to assess the internal consistency of the multiple item measures in the pre-questionnaire data. Using Cronbach's Alpha, multiple item measures were assessed ($n < 10$). The calculated Cronbach's Alpha values were all in the range of 0.67 and 0.89. Since the reliability was sufficiently high for all variables based on the guidelines by George Mallery [14], we used the means of these items as index measures. These values were calculated using SPSS in collaboration with another researcher, Jaap Dechering, who is also performing research on the acceptance of Sam.

2.4 Analysis methods

Qualitative analysis

The main form of data analysis performed during this research was an inductive thematic analysis applied to the free-text question mentioned in 2.2 [8]. An inductive thematic analysis is a form of analysis that generates themes inductively from qualitative data by coding the data. Inductive thematic analyses are used to identify themes that are strongly linked to the provided data [15]. This form of thematic analysis was appropriate, as the free-text responses that were looked at from the post-questionnaire are directly linked to this report's research question, therefore themes that are strongly linked to the provided data are desired.

The first step of the thematic analysis was to familiarize with the provided data. This was done by reading through all provided data, as well as studying the provided explanation files that accompanied the data files. Even though only the free-text question regarding why users found it easy or difficult to talk to Sam was used during the rest of the thematic analysis, all other free-text questions were also read during familiarization. It was important to become familiar with all aspects of the provided data and not just that single question to gain a proper understanding of the data set as a whole [16].

During this familiarization, unusable free-text responses were marked and unused for the rest of the analysis. This contained empty answers, non-relevant answers and answers that only referred to previous answers. There was a Spanish free-text response, which was translated to English in order to be coded. The response was not a complex one, therefore the chance of a misinterpretation or a wrong translation was minimal.

Investigator triangulation

Once familiar with the data, the coding step of thematic analysis was performed. First an initial list of possible codes was constructed. This list was extended throughout the first coding round in an iterative manner. Once this initial coding round was finished, a second round was done to ensure no codes were either missing or applied incorrectly. This resulted in the initial coded data and initial coding scheme, both of which can be find in the repository [10]. After these two rounds of individual coding, a second researcher, who also had a background in computer science, was asked to code the data with the initial coding scheme as part of the investigator triangulation [16]. This second researcher was first asked to study the initial coding scheme, after which they were shown how twenty responses should be coded. After this, they were asked to code twenty responses on which they received feedback. After this, the second researcher was deemed ready and was asked to code the remainder of the responses.

The Cohen's Kappa was calculated between the coding of the two researchers per each code. The code written and used was placed in the repository [10]. The first twenty responses were not taken into account, as these were used for training purposes. Any Cohen's Kappa below 0.6 is a weak agreement, whereas a Cohen's Kappa of 0.6 or higher indicates a moderate to good agreement [17]. Any codes that scored below 0.6 were either removed, reworded or merged with another code to increase the clarity and reproducibility of the coding scheme. Feedback from the second researcher was taken into account during this process.

The initial coding was then compared to the second researcher's coding and any conflicting codes were resolved through discussion. In the case that no agreement could be reached, a third researcher, who also had a background in computer science, was consulted to resolve the matter. Using the Cohen's Kappa values, feedback from the second researcher and discussions, the final coding scheme and final coded date was obtained. Both of these can be found in the repository [10]. Once the free-text responses from the post-questionnaire data were completely coded, initial themes were created by grouping codes.

Quantitative analysis

With these initial themes, the quantitative analysis was started, during which themes were updated in an iterative manner. These themes consisted of a number of codes, which indicated whether a response contained either zero or more themes. Rather than a binary number, a continuous number was given based on how many of the total codes of a theme a response had. If, for example, a response contained two out of four of the codes for a theme, that response would would have a value of 0.5 for that theme.

Hypotheses were made regarding possible correlations between user characteristics and the rating of the ease of talking to Sam, between user characteristics and themes or between themes and the rating of the ease of talking to Sam. The averaged data, as mentioned in 2.3 was also taken into account and was part of the user characteristics. Based on these hypotheses, any significant correlations were looked for using Pearson correlations. The software used to find and calculate these correlations can be found in the repository [10]. Correlations were deemed weak, moderate or strong based on the guidelines by Dancey and Reidy [18].

Method triangulation

To solidify any findings throughout the analysis, method triangulation was used [16]. This entailed that multiple data sources were used to attempt to find information that either backed up or contradicted any findings. This could, for example, be backing up findings from the qualitative analysis with ratings from the quantitative analysis, or backing up findings from the quantitative analysis with related literature.

3 Findings

Using the analysis methods as mentioned in 2.4, a total of six themes were identified in the free-text responses. Within these themes, two main aspects of talking to Sam were identified: actually using the application, e.g. clicking answer buttons or typing in answers manually, and how users felt about talking to Sam, e.g. how they felt about conversational agents in general or how at ease they felt whilst talking to Sam.

Some of these themes could indicate an aspect that requires improvements, whereas other themes could indicate qualities of the virtual coach that should remain unchanged. For each theme, recommendations were given regarding what to improve or what to leave unchanged.

The six themes that were identified are shown below in table 1. Each of these themes will be discussed separately later in this chapter. An overview of all themes, as well as their corresponding codes can be found in appendix B and in the repository [10]. Most notable here are t_3 and t_5, which were most frequent by both appearing 118 times. On average, users rated the ease of talking to Sam very high (M = 3.90, SD= 1.84).

ID	Theme	Occurrences
t_1	Conversations felt unnatural	42
t_2	Conversations felt natural	48
t_3	Clarity towards the user	118
t_4	Comfortable to talk	62
t_5	Ease of replying	118
t_6	Conversations felt impersonal	8

Table 1: An overview of all identified themes and their number of occurrences.

Multiple significant correlations were identified during the quantitative analysis that could provide an insight into what factors play a role in finding it easy or difficult to talk to the virtual coach. Table 2 shows an overview of correlations between each theme and the rating of the ease of talking to Sam. Furthermore, each theme's average rating was calculated.

Most notable here is the moderate, negative and significant correlation between the theme ”Conversations felt unnatural” and the ratings of ease of talking to Sam. A weak, positive and significant correlation between the theme ”Comfortable to talk” and the ratings was found. The other correlations that were significant were not strong enough to mean something (<0.10).

ID	Correlation	P-value	Average rating
t_1	-0.35	0.00	2.19
t_2	0.09	0.05	4.45
t_3	0.07	0.09	4.25
t_4	0.13	0.00	4.53
t_5	0.03	0.56	4.00
t_6	-0.09	0.05	2.38

Table 2: Correlation between themes and the rating of ease of talking to Sam.

Multiple correlations were looked for between certain user characteristics, ratings and the rate of ease of talking to Sam. More specifically age, openness to experience, the rating of the ease of the assigned activities, the rating of how motivated users were to do the assigned activities were looked at and the rating of the ease of talking to Sam were looked at.

Age was believed to possibly have an influence on the rating of the ease of talking to Sam, because elderly often have a negative attitude towards technology in general [19]. The ease and motivation users experienced during the assigned activities were believed to be important factors, because perceived usefulness is an important aspect in the user acceptance of information technology [20].

	Correlation	P-value
Age	-0.04	0.41
Openness to experience	0.04	0.40
Ease of assigned activities	0.30	0.00
Motivation during assigned activities	0.30	0.00

Table 3: Correlation between a number of characteristics and the rating of ease of talking to Sam.

As can be seen in table 3, two moderate and significant correlations were found. Both the perceived ease and motivation that users experienced whilst performing their assigned activities correlate with the rating of the ease of talking to Sam. This could possibly be explained by the fact that the perceived usefulness (in this case that of the assigned activities) is an important aspect in the acceptance of information technology.

3.1 Naturalness of conversation

In the free-text responses, many users mentioned that they found conversing with Sam to feel either natural or unnatural. To understand what this meant with regards to finding it easy or hard to talk to Sam, both of these themes were looked at in depth.

	t_1	t_2
Age (below average)	26	19
Age (above average)	22	23
Openness to Experiences (below average)	22	17
Openness to Experiences (above average)	26	25

Table 4: Number of responses with t_1 and t_2 based on age and openness to experience.

Table 4 shows the distribution of users who mentioned finding it either unnatural or natural to converse with Sam based on their age and openness to experience. The data does not show any notable differences and nothing could be concluded from it.

Conversations felt unnatural

> *”Clearly it is not possible to actually 'talk' properly with an online chat feature.” [11]*
>
> User P318

Notable in table 2 is the significant moderate negative correlation between t_1 and the ease of talking to Sam, combined with the low average rating of 2.19.

Though it might seem that, when conversations were perceived as unnatural, this would be something bad straight away, multiple users stated that they felt more comfortable with Sam than they would have been with a human. Reasons for this varied: some users mentioned that they felt they found the lack of human judgement comforting, other users stated they felt less pressure from a conversational agent, than they would have felt talking to a human. Literature suggests that consumers tend to prefer AI-based services over human-based services in contexts that require intimate interaction and the sharing of personal information.

> *”I felt uninhibited because I knew he wasn't a real person, so I didn't feel judged and it was easier to be honest.” [11]*
>
> User P107

Overall it seems that some people outright prefer an artificial conversation over a human alternative, whereas others straight away dislike it. Some users mentioned simply not liking the idea of artificial intelligence, whereas others stated that they did not see how artificial intelligence could truly understand and help them. Literature shows that this is a known phenomenon, as humans often seem to distrust AI due to their perceived lack of empathy [21].

People of older age often have a negative attitude towards technology in general [19]. Because of this, age was considered as a possible important factor to try and understand why some users prefer the artificial feel of Sam and why most users don't.

	Age (below average)	Age (above average)
Average rating	3.82	1.59

Table 5: The difference in ratings between below and above average aged users that found the conversation unnatural.

As shown above in table 5, a small difference between the above and below average age groups can be seen in the mean rating. An independent-samples t-test was conducted to compare the mean ratings in below mean age users and above mean age users. There was not a significant difference in the scores for below average age users (M=3.95, SD=1.69) and above average age users (M=3.83, SD=2.04); t(498)=0.73, p=0.46.

Conversations felt natural

> *"Talking with Sam felt very 'real' and I felt as though I was having a conversation with someone who was supportive and very friendly. I really liked the informal way of conversing also which I think added to the whole process feeling very friendly." [11]*
>
> User P229

Many users (48 responses) mentioned that they found conversing with Sam to feel natural. On the opposite side of the spectrum of t_1 is the significant, but weak, positive correlation between t_2 and ratings, combined with the higher average rating of 4.45. Part of these users mentioned also mentioned that they enjoyed Sam's human characteristics, such as being polite, friendly, understanding and empathetic. A number of users mentioned that because of the natural flow of the conversations, they found it easy to talk to Sam.

Overall, looking at average ratings, it seems that users who experienced conversing with Sam to feel natural (M=4.45) gave a higher rating than users who experienced it to feel unnatural (M=2.19). Many users reported that they enjoyed Sam's natural and informal language. They received a natural feeling, that was not trying to be too natural.

Recommendations

It is recommended that Sam's formulation and use of language does not change too much. Furthermore, it is often believed that increasing the naturalness of a conversational agent leads to better user experience [6]. Increasing this human-like way of speaking of Sam too much could in turn increase the uncanny perception some users were already getting. Currently, looking at the average ratings of the two themes, this balance should be shifted more towards natural. This must be kept in mind during further development of Sam.

To create a more natural perceived Sam, dynamic response times could be implemented as these are shown to lead to a more human-like perceiving of conversational agents [22]. To give users the feeling that Sam is able to express emotions to a certain degree, perhaps some form of visualisation could be added that shows this. For example, if Sam hears that the user is having a good time performing one of the assigned activities, a smiley could be shown.

3.2 Clarity towards the user

> *"He presented me clear choices and used simple phrases and words. It was easy to understand and talk to him." [11]*
>
> User P316

118 users stated in some form that they enjoyed the clarity of Sam towards them, with an average rating of 4.25 (SD = 1.64). Reasons for this varied from using clear and concise language to enjoying Sam's specific wording when asking questions, which left little room for misunderstanding. Users enjoyed knowing what was expected from them without having to put effort into trying to understand Sam. Misunderstandings of, for example, the activities that Sam proposed could lead to frustration, which is one of the big reasons that users stop using mHealth applications [6].

Albeit weak, a positive and significant (p=0.00) correlation of 0.15 was identified between clarity towards the user and the rating of the ease of talking to Sam. It seems that the reasons for finding Sam clear could be divided into two categories, users either say something about what Sam said or how it said it. Most frequently, users mentioned that Sam's phrasing was clear and straight forward. Users appreciate that Sam makes it clear what it wants or needs from the user and that it explains everything well, leaving little room for misunderstandings.

Recommendations

Sam's current use of language and wording seems to be enjoyed by a large portion of the users. The simple, informal and clear language used should remain unchanged and the way Sam phrases his requirements and questions is also very good. There is little to recommend, besides continuing this approach.

3.3 Comfortable to talk

> *"Sam was very polite and non-judgemental so I had no problems being open and honest with him." [11]*
>
> User P112

Users mention that Sam showed a variety of personality traits that made talking to Sam a comfortable experience. Users that mentioned that they it comfortable to talk to Sam, rated the ease of talking to Sam with a 4.53 on average. Users mentioned that Sam's friendly and polite way of talking made

Sam very likeable and made them feel at ease throughout the conversations. Users stated that they felt understood and that Sam showed empathetic behaviour, which was shown by Liu & Sundar to be favored by users talking to conversational agents over unemotional advice or support [23].

Moreover, it seems that part of the comfort users experienced, was related to the fact that the users were aware that Sam is not a human. User mentioned that, because they were aware that Sam was a bot, they believed that Sam was incapable of judging them or being impatient. Both of which users experienced as comforting thoughts. Currently, little research exists on improving user comfort when chatting with a conversational agent in mHealth.

Recommendations

Similarly to the theme "Conversations felt unnatural", it is recommended to keep the balance between human-like and bot-like, rather than to attempt to make Sam as human-like as possible. Furthermore, users experience writing their problems to conversational agents as comforting, even without receiving understanding behaviour from the conversational agent [24]. Therefore, adding in a question similar to "What problems did you face between our last meeting and today?" could allow users to freely express themselves, increasing user comfort. It is recommended to keep Sam's friendly, polite and understanding behaviour as this seems to be enjoyed by users.

3.4 Ease of replying

> *"I loved that there were a wide range of fixed responses to most of the questions, this made it very easy." [11]*
>
> User P274

Users that mentioned finding it easy to reply to Sam, rated the ease of talking to Sam with a 4.00 on average. Conversation with the virtual coach mostly consists of replying to Sam. Therefore the ease of replying greatly impacts the ease of use of the application in general. According to Venkatesh et al. [20] the ease of use is one of the factors that play a role in the individual acceptance of information technology. Therefore it does make sense that the ease of replying would affect the user acceptance of the application.

Whenever users wanted to reply something that was not provided in one of the provided answers buttons, they were able to respond using free-text. User experiences with these free-text replies were mixed: five people mentioned that Sam was able to handle these free-text answers well, but three people mentioned that Sam was unable to properly handle and reply to their free-text messages. Sam being unable to understand free-text messages could lead to frustrating scenarios for users.

Recommendations

In general, users seem to enjoy the provided answer buttons very much. It is recommended to keep these answer buttons and possibly extend these to provide a broader range of answers. If possible, further analysis could be performed on the free-text answers that users have given to derive what type of answers should be added to the provided buttons. Some users seemed to refrain from using free-text replies, as Sam was unable to process them well. It is recommended to perform additional research and to improve this.

3.5 Conversations felt impersonal

> *"The flow of the conversation was pretty good, but it felt somewhat standardised. This is to be expected from a conversation with a bot I would say. Ideally, I'd like to be able to say something more about my current situation, what my goals were etc. and just have it more personalised." [11]*
>
> User P481

Though only eight users mentioned it explicitly, a number of users stated indirectly that the conversations felt restrictive because they could not ask Sam anything. A number of these users stated that the conversations felt impersonal due to this. Another number of users stated that they did not get to know Sam and that it felt like Sam was too straight to the point. Users seem to expect a certain level of interactivity, based on Sam's human-like features and therefore it is important to satisfy these expectations to a certain degree.

Recommendations

According to Go & Sundar [25], a high level level of message interaction can, in some cases, compensate for the impersonal nature of conversational agents. As implementing a full two-way conversation is a very expensive and difficult task, it is recommended to opt for another way to make the conversations feel more personal. Moreover, a full-scale two-way conversation could have negative effects on the effectiveness of the application, as it becomes easier to stray off topic. Making conversing Sam feel more personal could be achieved in a number of alternative approaches. It is recommended to give Sam more of a personality and to have it give some information about itself. This could allow users to feel like they get to know their virtual coach. Another way this could be achieved is by allowing Sam to tailor to their users over time. Sam could remember show users that it remembers things that they said by mentioning them later on for example.

4 Responsible Research

Throughout this research, multiple measures have been taken to provide responsible research. First of all, during every step, reproducibility was taken into account. To maximize the reproducibility of this research, all necessary steps that were taken have been stated in this report. Furthermore, all software that was used, was written during this research and published on GitHub along with all interim results [10]. The GitHub repository contains a "readme.txt" in which is explained how the written software should be used or which files contain which information. The original data-set was also publicly available [11].

The problem, however, is that some steps during the analysis of qualitative data are less reproducible due to the researcher's bias. To minimize this bias, two fellow researchers, Mahira Ali and Nadyne Aretz, were involved in this research as mentioned in 2.4 as part of the investigator triangulation.

Furthermore, use was made of method triangulation to solidify any findings. This lowers the researcher's bias, as multiple sources of data are taken into account [16]. Furthermore, during the quantitative analysis measures were taken to avoid p-hacking by only looking for correlations based on hypotheses [26].

Lastly, this research was not performed without guidance of peers and supervisors. Through mutual meetings with the supervisors of this research and through conceptual discussions amongst peers, this research was able to receive feedback from multiple sources. The four fellow researchers that also performed research on the acceptance of Sam are thanked and named here as a form of acknowledgement: Nadyne Aretz, Omar Sheasha, Mahira Ali and Jaap Dechering.

5 Discussion and Conclusion

This research's goal was to give recommendations to improve Sam by identifying the eases and difficulties in talking to it. Through analysis of free-text data in the post-questionnaire, filled in by users, a total of 6 themes were identified using an inductive thematic analysis. Conversations felt unnatural, conversations felt natural, clarity towards the user, comfortable to talk, ease of replying and conversations felt impersonal. Moreover, the eases and difficulties talking to Sam could be divided into two main categories: finding it easy or difficult to talk to Sam linguistically (1) and emotionally (2).

The majority of users were positive about Sam and rated the ease of talking to it high on average (M = 3.90, SD= 1.84). Though there still was some room for improvement, part of finding it easy or difficult to talk to Sam will always be subjective. Based on the identified themes, recommendations were given in order to make it easier for users to talk to Sam. Furthermore, a correlation between the rating of the ease of talking to Sam and both the ease and the motivation users experienced during the assigned activities was found. This might be explained by the perceived usefulness of talking to Sam having an influence on the acceptance of it.

Naturalness of conversation: There seems to be a delicate balance between being too artificial and being too human-like for Sam. Currently, users think that Sam is not shifting too much to one side of this balance and they enjoy Sam's natural and informal way of speaking. This balance should be kept in mind during any future developments. Sam could still benefit from being perceived as a bit more natural, which could be achieved by implementing dynamic response times and some form of emotion visualisation.

Clarity towards the user: Currently, Sam has a very clear and simple way of speaking. Furthermore, Sam thoroughly explains his messages and makes it clear what it requires from users. Users seemed to enjoy this, as this left little room for misinterpretations. It is recommended that this clarity remains unchanged during any future developments.

Comfortable to talk: Users find Sam to be comfortable to talk to. Sam is perceived as a friendly, understanding and empathetic virtual coach. As an improvement, Sam could add a question where it asks for any problems users have faced, allowing for users to express themselves freely. Simply writing these thoughts to a conversational agent has been proven to be comforting, even without a response from Sam [24].

Ease of replying: In general, users found it easy to reply to Sam using the provided answer buttons. Therefore it is recommended to keep these provided answer buttons, possibly even extending them to cover more answers. Opinions of Sam's processing of free-text replies were divided, meaning that there is still room for improvement. It is recommended to perform further analysis of these free-text responses and identify cases that Sam was unable to process correctly.

Conversations felt impersonal: Some users mentioned that the interactions with Sam felt impersonal. To improve this, it is recommended to give Sam a very basic personality and background, creating more of a character that users can get to know. Users could either ask Sam about itself or Sam could tell things about himself. Great care must be taken to not humanize Sam too much doing this, as this might lead to an uncanny feeling for users [27]. To not create a too human-like personality, these things could be as simple as Sam telling users how long he has been a virtual coach or how many users it has talked to. Furthermore, it is recommended that Sam remembers things users said. This would show that Sam somewhat personalizes over time and is not standardized to most users. As long as care is taken into not having any bias or unethical views in Sam's simple personality and background, this should not lead to any ethical issues. Though there is still room for improvement of Sam, these recommendations should form a good start to improving talking to Sam.

Unfortunately, this study had its limitations. Throughout this research, a variety of measures were taken to reduce the researchers' biases and to make this research as reproducible as possible. However, it is impossible to completely rid this research of all bias. Due to time constraints, the research was not as elaborate as it could have been. For example, the investigator triangulation could have been more elaborate, reducing bias even further. Having too few double coders could be a limitation for the reliability of the conclusions. Moreover, different types of correlation could have been taken into account, instead of just Pearson's. This limited identifying additional themes or explore the existing themes more. Furthermore, additional research could've been performed into causation, instead of only looking at correlations, possibly providing more insights into the causes of the identified themes. This could have allowed for more pinpointing of specific reasons to find it easy/difficult to talk.

6 Future work

This research intended to provide recommendations to improve talking to Sam. In the future, more observational studies should be performed in the future to assess the effectiveness of these recommendations.

The main question for this research was quite broad: talking to Sam could mean a couple of things. This could be regarding actual usage of the application, how easy and hard is it to send a reply for example. It could also mean how easy or hard it mentally was for people to talk to Sam. In the future, more precise research questions could be formulated and looked into to provide more accurate answers. For example, ”What are reasons to find it easy or difficult emotionally to talk to Sam?".

As mentioned in 3.4, the currently provided button answers could perhaps be extended using analysis of the free-text replies users give. During this analysis, the issue where some free-text replies were not recognized by Sam could also be looked into. Identifying the types of answers that users frequently give using free-text might give an indication of what types of answers are missing from the provided answer buttons.

References

[1] J. Bergström and H. Preber, "Tobacco use as a risk factor," *Journal of periodontology*, vol. 65, pp. 545–550, 1994.

[2] W. H. Organization *et al.*, *WHO global report on trends in prevalence of tobacco use 2000-2025*. World Health Organization, 2019.

[3] M. S. Marcolino, J. A. Q. Oliveira, M. D'Agostino, A. L. Ribeiro, M. B. M. Alkmim, and D. Novillo-Ortiz, "The impact of mhealth interventions: Systematic review of systematic reviews," *JMIR mHealth and uHealth*, vol. 6, no. 1, e8873, 2018.

[4] J. B. Bricker, K. E. Mull, J. A. Kientz, *et al.*, "Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy," *Drug and alcohol dependence*, vol. 143, pp. 87–94, 2014.

[5] A. S. Mustafa, N. Ali, J. S. Dhillon, G. Alkawsi, Y. Baashar, *et al.*, "User engagement and abandonment of mhealth: A cross-sectional survey," in *Healthcare*, Multidisciplinary Digital Publishing Institute, vol. 10, 2022, p. 221.

[6] E. Deursen, "The paradoxical impact of mobile health applications: Why people stop using them.," 2021.

[7] N. Albers and W.-P. Brinkman, "Perfect fit - experiment to gather data for and test a reinforcement learning-approach for motivating people," May 2021. DOI: 10.17605/OSF.IO/K2UAC. [Online]. Available: osf.io/k2uac.

[8] M. E. Kiger and L. Varpio, "Thematic analysis of qualitative data: Amee guide no. 131," *Medical teacher*, vol. 42, no. 8, pp. 846–854, 2020.

[9] R. Nancy Carter, D. Bryant-Lukosius, and R. Alba DiCenso, "The use of triangulation in qualitative research," in *Oncology nursing forum*, Oncology Nursing Society, vol. 41, 2014, p. 545.

[10] A. Ekinci, *Analysis of the ease and difficulty of talking to the virtual coach.* Version v1.4, If you use this software and/or data, please cite it as below., Jun. 2022. DOI: 10.5281/zenodo.6647544. [Online]. Available: https://doi.org/10.5281/zenodo.6647544.

[11] N. Albers, M. A. Neerincx, and W.-P. Brinkman, "Acceptance of a Virtual Coach for Quitting Smoking and Becoming Physically Active: Dataset," May 2022. DOI: 10.4121/19934783.v1.

[12] N. Albers and W.-P. Brinkman, "Acceptance of a virtual coach for quitting smoking and becoming more physically active: A thematic analysis," *Interactive Intelligence - TU Delft*, 2022.

[13] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[14] D. George and P. Mallery, *Spss for windows step by step: A simple guide and reference. 11.0 update , 2003*, 2016.

[15] M. Javadi, K. Zarea, *et al.*, "Understanding thematic analysis and its pitfall," *Demo*, vol. 1, no. 1, pp. 33–39, 2016.

[16] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[17] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[18] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson education, 2007.

[19] B. Knowles and V. L. Hanson, "The wisdom of older technology (non) users," *Communications of the ACM*, vol. 61, no. 3, pp. 72–77, 2018.

[20] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[21] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, "In ai we trust? perceptions about automated decision-making by artificial intelligence," *AI & SOCIETY*, vol. 35, no. 3, pp. 611–623, 2020.

[22] U. Gnewuch, S. Morana, M. T. Adam, and A. Maedche, "Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction," in *26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank*, 2018, p. 143 975.

[23] B. Liu and S. S. Sundar, "Should machines express sympathy and empathy? experiments with a health advice chatbot," *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 10, pp. 625–636, 2018.

[24] L. Medeiros, T. Bosse, and C. Gerritsen, "Can a chatbot comfort humans? studying the impact of a supportive chatbot on users' self-perceived stress," *IEEE Transactions on Human-Machine Systems*, 2021.

[25] E. Go and S. S. Sundar, "Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions," *Computers in Human Behavior*, vol. 97, pp. 304–316, 2019.

[26] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLoS biology*, vol. 13, no. 3, e1002106, 2015.

[27] C. M. van Lierop, “The uncanny valley: The paradoxical effects of enhancing the human-likeness of customer support chatbots on perceived trust and satisfaction,” 2021.

A Final coding scheme

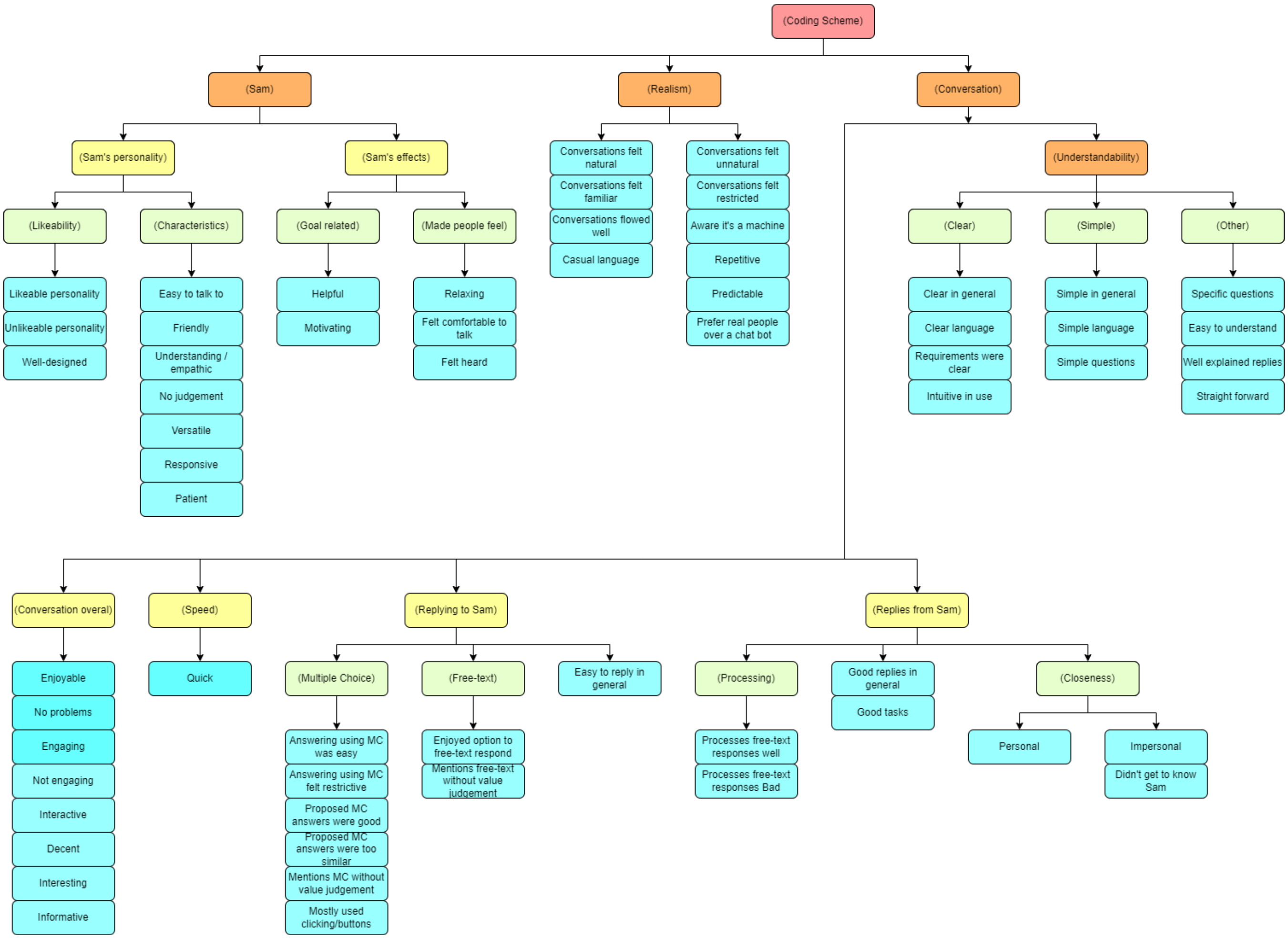

Figure 1: The final coding scheme.

B Final themes

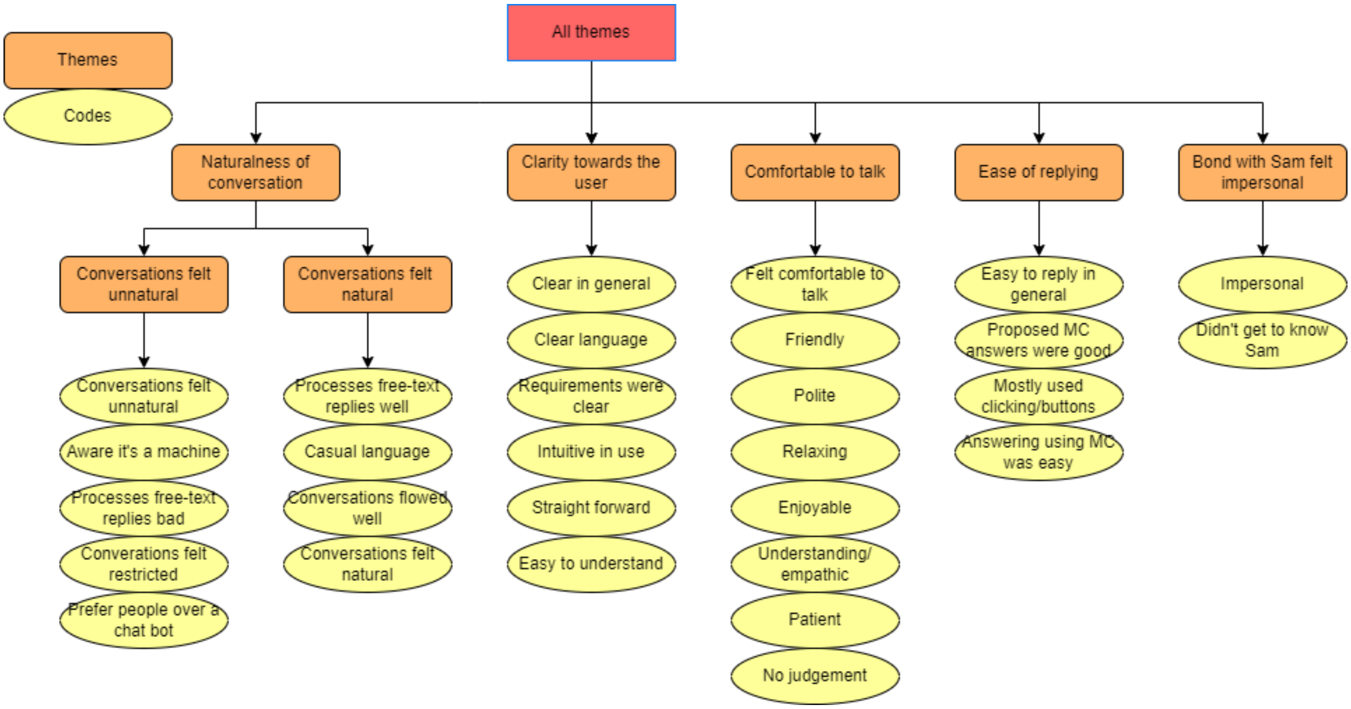
Themes
Codes
All themes
Naturalness of conversation
Clarity towards the user
Comfortable to talk
Ease of replying
Bond with Sam felt impersonal
Conversations felt unnatural
Conversations felt natural
Conversations felt unnatural
Aware it's a machine
Processes free-text replies bad
Converations felt restricted
Prefer people over a chat bot
Processes free-text replies well
Casual language
Conversations flowed well
Conversations felt natural
Clear in general
Clear language
Requirements were clear
Intuitive in use
Straight forward
Easy to understand
Felt comfortable to talk
Friendly
Polite
Relaxing
Enjoyable
Understanding/ empathic
Patient
No judgement
Easy to reply in general
Proposed MC answers were good
Mostly used clicking/buttons
Answering using MC was easy
Impersonal
Didn't get to know Sam

Figure 2: An overview of all themes and their corresponding codes.