03/02/2021

Testing a scale for perceived usability and user satisfaction in chatbots: Testing the BotScale

Master thesis

Mirre van den Bos
s1806025

First supervisor: dr. Borsci

Second supervisor: prof. dr. van der Velde

University of Twente
BMS Faculty
Department of Psychology

Abstract

Currently, a user satisfaction scale for chatbots does not exist, while more and more companies may start using and developing chatbots. The present study aimed to replicate the study by Balaji and Borsci (2019), who have developed a 42-item chatbot scale (BotScale) with four factors, by also proposing a reduced the BotScale to 17 items, with the same number of factors.

A replication was done by involving fifty volunteers in the assessment of nine chatbots by using an English and Dutch version of the BotScale. Additionally, an already existing scale for testing user satisfaction in voice interfaces, the Speech User Interface Service Quality (SUISQ-R) (Lewis & Hardzinski, 2015), was used to check for external validity.

A principal component analysis on a chatbot x item dataset was performed to reduce the original scale from 42 to 14 items by looking at factor loadings, item-total correlations and reliability. Results show that a four-factor structure for the BotScale captures 85% present of variance and the 14-item BotScale had a reliability of $\alpha = 0.93$. A correlational analysis between the Dutch and English version of the BotScale showed that the Dutch translation was as reliable ($\alpha = 0.89$) as the original scale ($\alpha = 0.93$). Furthermore, the Dutch version had a significant strong positive correlation to the English version of the BotScale. Finally, a correlation between the BotScale and the SUISQ-R was performed, and results suggested a significant positive moderate correlation. However, when looking at correlations between factors, it seems that the SUISQ-R does not contain all the important aspects to evaluate chatbots.

Keywords: Chatbots, user satisfaction, usability, BotScale, SUISQ-R, voice interface.

Acknowledgements

First of all, I owe my gratitude to dr. Simone Borsci for his guidance the past year. But most of all I want to thank you for your kind words when I needed motivation. Sometimes the pandemic made it hard to continue working in the normal pace, but you have kept me motivated through this whole thesis. I also want to thank prof. dr. Frank van der Velde for his feedback and different perspective which helped elevating the level of the thesis.

I also want to thank Nando for his help on all the different occasions. Thank you for proofreading my thesis, even though Psychology is not your field of expertise. I also want to thank you for listening to me whenever I had full discussions with myself about which choices I had to make and why. I also want to thank my parents for always believing in me, even if you have no clue what I am actually working on. I also want to thank my brother, for helping me think about other things than studying.

I also want to thank my friends and roommates. I want to thank my roommates for helping me whenever they can, brainstorming with me and creating a space where I can be myself. I want to thank Stef for giving me advice about research related topics, and in general for all the great moments we had during our study. I want to thank my friends Suzanne, Miriam, Kim and Eline for the time we had together during my time here in Enschede. I also want to thank everyone at Chassé for motivating me and giving me a scheduled time to take a break.

Lastly, I want to thank everyone who has participated, without you I literally could not have finished this thesis.

Contents

1. Introduction

1.1 Chatbots and their uses

Nowadays, you never know for sure when you are chatting to an actual person when browsing the web or using customer service. Sometimes the responses of the customer service agent are slow, or a bit out of context which makes you think you are not actually talking to a human. It could be that you are talking to a chatbot, which is a virtual agent that can interactively talk to humans through natural language (Przegalinska, Ciechanowski, Stroz, Gloor, & Mazurek, 2019). Bavaresco et al. (2020) performed a systematic literature review to examine in which fields chatbots are utilised and for what goal they are used. About 40% of the included studies of this research are within the commerce domain and chatbots are most commonly used for Q&A and customer support (Bavaresco et al., 2020). Currently, only around 14% of Dutch commercial and non-commercial organisations use a chatbot, however 47% of these organisations plan to integrate a chatbot in the coming two years (van Os, Hachmang, Akpinar, Kreuning, & Derksen, 2018). Even though a small percentage of Dutch companies currently use a chatbot, it seems there is an interest from more companies to also start using a chatbot. Therefore, it is important that these chatbots are up to a standard where the consumers feel that they are easy to use and experience their added value. However, according to a bibliometric analysis by Io and Lee (2017), not a lot of research has been done about the interaction between humans and chatbots. Therefore, the aim of this study is to test a scale with which chatbots can be evaluated on their usability so they can be improved. Which should improve the interaction between humans and chatbots.

To be able to understand how chatbots and humans interact, it is also important to know what a chatbot entails. There are different kinds of chatbots, with different goals and different ways of interacting with humans. Adamopoulou and Moussiades (2020) summarised all the different categories in which chatbots can be divided. One way to differentiate chatbots is by the underlying development techniques, e.g. using machine learning or pattern-matching. A chatbot can also be categorised based on how they communicate to the user, which could be via text, an artificial voice or even with images. As mentioned before, chatbots can also have different goals, from giving information, to performing tasks or just for having conversations with the user. Furthermore, chatbots

can also be divided in what manner they interact, they can help without having friendly interactions (interpersonal), they can act like a companion (intrapersonal) or chatbots can even interact with each other (inter-agent). Another possible way to divide chatbots into categories is by looking if the chatbots performs all tasks by itself, or whether encounters problems or the task is beyond its capabilities asks for a human agent to take over (Adamopoulou & Moussiades, 2020).

In research, several possible factors which influence the interaction between chatbots and humans have been found. Firstly, several studies look at the humanisation of chatbots (Go & Sundar, 2019; Jenkins, Churchill, Cox, & Smith, 2007; Qiu & Benbasat, 2009). Qiu and Benbasat (2009) aimed to study the influence of anthropomorphic cues on social presence of product recommendation agents and consequently how social presence influences the user to keep using the agent. They performed a laboratory experiment and tested the effects of showing a face with the agent and either using text, text-to-speech or a human voice to communicate. Their results show that anthropomorphic cues significantly affect social presence which affected trust and perceived enjoyment. These two then affected perceived usefulness which affected usage intentions. When comparing human speech, text-to-speech and text to each other, human speech had the most positive influence on social presence. There was no difference between text-to-speech and text and their influence on social presence (Qiu & Benbasat, 2009).

Go and Sundar (2019) also studied this humanisation of chatbots. They looked at visual cues (human face or an icon), identity cues (human or chatbot) and message interactivity. Cues could either be high or low on human alikeness, resulting in a 2 x 2 x 2 between-subject design. They found a few results which can be useful when designing a chatbot. First, they found that if participants knew the chatbot was a chatbot, they were more satisfied then when they thought it was human. Second, higher message interactivity caused higher satisfaction and social presence. The same holds for the visual cues. Lastly, high message interactivity could compensate for low visual cue and vice versa (Go & Sundar, 2019).

Another factor which could possibly play a role is the productivity of the chatbot, intended as how quickly or effectively they can help the user. Productivity seems rather obvious since you would probably only use a chatbot when you perceive the help of the chatbot to be faster than doing a task by

yourself. Zamora (2017) studied what people expect from chatbots by letting participants perform tasks with a chatbot and afterwards answer questions. Participants did mention that a chatbot should save them time, which they now felt was not always the case. Additionally, Ben Mimoun, Poncin, and Garnier (2017) used eye-tracking to see where participants look when interacting with a chatbot. They found that participants barely looked at the anthropomorphic features of the chatbot. Participants looked at the text where communication with the chatbot took place. They state that the visual cues may not be as important as for example productivity. It was also found that when using the animated conversational agent, efficiency improved. However, the degree to which the chatbot could be used by participants only had an influence on the objectively measured efficiency but not on the perceived efficiency by the participant. Their second study looks at individual differences. They found that higher internet skills lead to finding the chatbot more useful and needing less time and effort to use the chatbot (Ben Mimoun et al., 2017).

Another important factor is trust, since it may determine the amount of information we are willing to share. Zamora (2017) also found that some participants stated that some topics or information is too personal to give to a chatbot. These concerns stem from fear the data will be mishandled or leaked. However, some participants would like to discuss private matters with a chatbot, since they will not be judged which can save embarrassment. For this to work, there needs to be trust (Zamora, 2017).

Wang and Benbasat (2008) aimed to identify how to build trust in agents. In their study, they used a product recommendation agent within the e-commerce domain. Their results show that there are four ways in which trust can be built. Trust can be built by gaining information about the agent and by considering potential loss or rewards when trust is put in the agent. It can also be built by experiencing the interaction with the agent and it can be based on predispositions in early phases. The predispositions which are meant here, is the tendency to trust the technology or not based on your previous experiences or your attitude. All these mentioned factors can have a positive influence on trust, however the factors about interaction and considering loss or rewards could also have a negative influence. An upside is that positive reasons had more influence than negative ones. They also found that heuristic cues and institutional rules do not contribute to trust, positively or negatively. Lastly,

experiences are very important in building trust and this is also why trust is quite sturdy (Wang & Benbasat, 2008). Additionally, a study by Waytz, Heafner, and Epley (2014) showed that anthropomorphism in automated cars led to higher trust, which could possibly also be the case for chatbots.

In sum, a few main characteristics which are important form human-chatbot interaction were humanisation of chatbots, productivity and trust of the user in the chatbot. These characteristics could all help in trying to develop a user-friendly chatbot, however it is also important to investigate how this user-friendliness or satisfaction can be measured.

1.2 Measurement of chatbot satisfaction

Botanalytics (2018) surveyed different companies to find how analytics about engagement with chatbots can help companies to improve on them. However, they found that the interpretation of analytics about conversation length or retention rate, can differ per domain. For example, they mention a short conversation length is good in the financial sector, since this means the customer probably got the correct answer quickly. However, in the entertainment sector you want a long conversation length since this means the customer is being entertained and wants to keep talking (Botanalytics, 2018). However, these metrics do not give a deeper insight into actual user satisfaction and do not show how user satisfaction could be measured.

Instead, user satisfaction can be used to improve chatbot, however currently there is not a user satisfaction scale specifically designed for chatbots. Therefore, researchers are using different methods to assess user satisfaction. Some researchers use several existing scales for modern technology together and sometimes also adapt them (Araujo, 2018; Chung, Ko, Joung, & Kim, 2020; Sajjadi, Hoffmann, Cimiano, & Kopp, 2019; Sheehan, Jin, & Gottlieb, 2020), some create their own items based on other research (Chung et al., 2020; Lee & Choi, 2017) while others use different measures than user satisfaction to gain insights (Schuetzler, Giboney, Grimes, & Nunamaker, 2018). However, this means that chatbots are not always assessed in a similar way. For example, Sheehan et al. (2020) looked at perceived usefulness, perceived ease of use and adaptation rate, while Chung et al. (2020) took interaction, entertainment, communication quality and overall satisfaction into account, and again

Lee and Choi (2017) investigated self-disclosure, reciprocity, trust, interactional enjoyment and user satisfaction.

Another possibility is to use an existing user satisfaction scale not specifically made for chatbots. Three well known standardised usability scales were analysed and compared by Borsci, Federici, Bacci, Gnaldi, and Bartolucci (2015). They analysed the System Usability Scale (SUS) (Brooke, 1996), the Usability Metric for User Experience (UMUX) (Finstad, 2010) and the UMUX-lite (Lewis, Utesch, & Maher, 2013). Borsci et al. (2015) compared the different scales to each other and how they performed under different amounts of time the user could interact with the product. In the end, all scales correlated strongly with each other, even on the different conditions. However, these scales are not specifically made for assessing user satisfaction in chatbots and Tariverdiyeva and Borsci (2019) found that the UMUX-lite does not cover all factors which were found to be important in chatbots. Since Borsci et al. (2015) found these three standardised scales to find similar results, it could then be assumed the other scales also do not cover all the important factors. Additionally, a scale including all possible important aspects could also give more detailed feedback to designers to tackle.

Next to looking at general user satisfaction scales, scales developed to programs related to chatbots could also be considered. Scales have been developed for voice interfaces, which share some properties with chatbots. Voice interfaces can also interact with humans, but do so via natural speech instead of text (Interaction Design Foundation, n.d.). Three examples of these scales are the Mean Opinion Scale (MOS-X) (Lewis, 2018) , the Subjective Assessment of Speech System Interface (SASSI) (Hone & Graham, 2000) and the Speech User Interface Service (SUISQ) (Polkosky, 2005) , which were compared by Lewis and Sauro (2020). An overview of these scales and their items can be seen in Appendix A.

Over the years, the MOS has been redeveloped into the MOS-X by Polkosky and Lewis (2003). Some properties of the MOS-X are:

- MOS-X has 15 items (Polkosky & Lewis, 2003).
- Has a shorter version with 4 items, the MOS-X2 (Lewis & Sauro, 2020).
- Items ask primarily about the voice of the system (Lewis & Sauro, 2020).

Secondly, the SASSI is another scale for voice interfaces developed by Hone and Graham (2000) with the following attributes:

- Has 34 items (Hone & Graham, 2000).
- Items ask about the whole interface, not only the voice, including accuracy, usefulness, annoyance and more (Hone & Graham, 2000).
- Only construct validity is assessed in the study of Hone and Graham (2000), no analysis was done for concurrent validity, overall reliability and sensitivity (Lewis & Sauro, 2020).

Lastly, the SUISQ was developed by Polkosky (2005) and later was updated and reduced by Lewis and Hardzinski (2015) to the SUISQ-R and SUISQ-MR. Some of its properties are:

- SUISQ has 25 items and four factors (Polkosky, 2005).
- SUISQ-R has 14 items and four factors and is still reliable (Lewis & Hardzinski, 2015).
- SUISQ-MR has 9 items and four factors, however it is recommended to only use this when it is really needed (Lewis & Hardzinski, 2015).
- Items include the voice, timing and if users want to keep using the interface (Lewis & Hardzinski, 2015).

A preliminary version of a scale to evaluate user satisfaction in chatbots was developed by Balaji and Borsci (2019), based on a systematic literature review and a usability test by Tariverdiyeva and Borsci (2019). Balaji and Borsci (2019) followed this by doing another systematic literature review and asked experts to review the factors found so far which seemed important when evaluating a chatbot. Items were developed and a focus group was held to get feedback on the items and factors. The feedback was incorporated, and the item pool was once more tested by letting participants interact with chatbots and afterwards answer the items from the item pool. Their original version included 42 items and their shortened version included 17 items, with a four-factor structure. Additionally, Silderhuis and Borsci (2020) aimed to replicate the study by Balaji and Borsci (2019), while also comparing age groups.

1.3 Present study and aims

The present study aims to replicate the study from Balaji and Borsci (2019). In their research, they mention a few limitations, which is why further research is needed. The first research question will investigate if it is possible to achieve a similar factor structure to Balaji and Borsci (2019), who found a four-factor structure. One of their limitations stated that they did a participant x item analysis, but possibly a chatbot x item analysis will yield a different structure. Therefore, the first research question will be:

- RQ1: Is the factorial structure of the questionnaire in line with previous studies?

Moreover, Balaji and Borsci (2019) and Silderhuis and Borsci (2020) both reduced the Bot Usability Scale into a more comprehensive scale which could then also reduce strain on the users. Therefore, this study will also try to reduce the scale using the same kind of factorization and then compare it to the reduced scales of previous studies. A shorter scale can reduce the strain on users and remove redundant items. Thus, the second research question is:

- RQ2: Can the BotScale be shortened, while keeping high reliability and including items about every chatbot feature?

Additionally, a Dutch version of the BotScale has been previously developed. Therefore, a second aim is to compare the Dutch and English version of the BotScale to see if the Dutch version is translated correctly while keeping the intended meaning of the statements. It is important to test the translation since it is not always possible to have a completely similar translation while keeping the intended meaning. Additionally, translating this scale to another language, in this case Dutch, may improve the user experience while filling in the scale for Dutch users. If the chatbot that the user is evaluating also uses the Dutch language, answering the questions in Dutch may yield more reliable results than a non-native language scale. Which results in the following research question:

- RQ3: Does the Dutch translation of the BotScale correlate with the original version?

Lastly, since scales for voice interfaces are used instead of general user satisfaction scales, this shows there could be a need for more specialised scales. This study will also test to see if a scale for

voice interfaces could be applied to evaluate a chatbot which uses text to interact. There is currently no scale available to assess chatbots, however there are scales which assess voice interfaces (Hone & Graham, 2000; Lewis, 2018; Lewis & Hardzinski, 2015). Voice interfaces are similar to chatbots in that they both use a form of language to interact with the user. When initially looking at items of voice interface scales, it can be seen that some items directly ask about the voice (Lewis, 2018; Lewis & Hardzinski, 2015), which would not be applicable to the only-text chatbots. Comparing a scale specifically designed for a chatbot and a scale designed for a voice interface could show if there are important differences when evaluating chatbots. The MOS-X only evaluated the voice itself (Lewis, 2018), thus using this scale would not be relevant since the chatbots used in this study only use text to interact. The SASSI has not been updated in twenty years and lacks verification of overall reliability (Hone & Graham, 2000). The SUISQ assesses the overall usability of a voice interface and has tested the scales on reliability, validity, etc. Additionally, a shorter version is also available (Lewis & Hardzinski, 2015), which can decrease the strain on the participant during the study. Therefore, another aim is to look at the relationship between the BotScale and the SUISQ-R. Since there is no Dutch version available of the SUISQ-R, it has to be translated since participants will also be able to answer with the Dutch BotScale. It is not an aim of this study to validate a Dutch version of the SUISQ-R, but reliability will be checked to see if it can be used to compare the Dutch SUISQ-R to the Dutch BotScale. The last research question therefore only focusses on the comparison of the SUISQ-R and the BotScale:

- RQ4: Do the BotScale and SUISQ-R show a moderate to strong correlation when comparing the ratings of the chatbots?

2. Methods

2.1 Participants

Through convenience sampling, 50 volunteers participated in the present study (*M*age = 23.32, *SD*age = 4.04). About 34% of participants were male and 66% was female and 82% of participants were Dutch, 12% were German and 6% had other nationalities. Participants could choose whether they wanted to answer the English or Dutch version of the questionnaire. 4% of participants were extremely familiar with chatbots, 24% very familiar, 38% moderately familiar, 30% slightly familiar and 4% not familiar at all. Furthermore, 88% had definitely or probably used a chatbot and the rest were unsure or hand never encountered a chatbot. Lastly, participants were asked how often they use a chatbot when they answered on the previous question that they had used a chatbot before, to which 76% answered never, 6% rarely and 6% daily or a few times a week. The research was approved by the Ethics Committee of the BMS faculty of the University of Twente. Before participating, participants read an information sheet and agreed with the informed consent (See Appendix B). Additionally, Psychology and Communication Science students from the University of Twente could earn course credits if they signed up through the corresponding system.

2.2 Materials

Qualtrics (n.d.) was used to gather data using an online questionnaire. Participants received an anonymous link after they entered the online meeting. Within Qualtrics (n.d.), the developed 42-item BotScale (See Appendix C) from Balaji and Borsci (2019) and the SUISQ-R (Lewis & Hardzinski, 2015) were presented after each interaction with a chatbot. The BotScale uses a 5 point Likert scale and orginally the SUISQ-R uses a seven-point Likert scale, so it was chosen to also use the SUISQ-R with a five-point Likert scale. Additionally, a Dutch version of the BotScale (See Appendix D) from Silderhuis and Borsci (2020) was used and the SUISQ-R was translated to Dutch (See Appendix E). Lastly, the chatbots and tasks were also presented in Qualtrics (n.d.). Most of the tasks were similar to the tasks used by Balaji and Borsci (2019) and were also translated to Dutch (See Appendix F). Due to the Covid-19 pandemic, Google Meet. (n.d.) was used to have online meetings with the participants.

2.3 Task

The tasks consisted of finding and using five out of nine chatbots and answering the two scales afterwards. First, the participant had to read the scenario for which they would be using the chatbot (See Appendix F). Subsequently, they had to copy the link of the website and find the chatbot themselves. Once they had found the chatbot, they could ask their questions until they received an answer they were satisfied with. After this, they could go on to the questions. After the completion of the task with every chatbot, participants had to fill out the two scales. For both scales the items were randomised each time and the two scales were also randomised in order.

2.4 Procedure

Participants received a link to Google Meet. (n.d.), ten to fifteen minutes before the start of the online session. When the participants entered the meeting, the researcher would share the Qualtrics (n.d.) link and explain the main goal and the difference between the English and Dutch version. After choosing the language, participants could read an information sheet and after that had to actively sign the informed consent. If they ticked yes on the question about recording the session, the researcher would start the recording. If they agreed to the rest of the informed consent, they would continue to the demographic questions and questions asking about their previous chatbot experiences. Before continuing to the tasks, the researcher would explain how the task would go. Additionally, the researcher pointed out that participants were not tested on how skilled they are with chatbots, but that participants would test the chatbot to be able to evaluate them in the scales. After the explanation, participants could continue at their own pace and perform the task with each of the five chatbots. When the task was completed, they had to fill out the 42-item BotScale (Balaji & Borsci, 2019) and the SUISQ-R (Lewis & Hardzinski, 2015). Meanwhile, the researcher stayed in the session, so that any questions or insecurities could be answered. When the participant was finished with all the tasks, the researcher would ask if they had questions, if it went well and lastly would thank them for their participation.

2.5 Data Analysis

Data was exported out of Qualtrics (n.d.) to Excel in numeric values. In Excel, unnecessary columns of data were deleted, labels were given to the items and the data was rearranged so it could be imported into R (v4.0.2; R Core Team., 2020). The dataset was composed of data generated with the Dutch and English version of the scale. Balaji and Borsci (2019) suggested that analyses on the structure of the scale could be explored with a classic psychometric approach (participant x item) and with tables organised per chatbots (chatbot x item) in order to look at the differences in the answers for each chatbot, instead of how each participant reacted to each design.

To answer the first research question concerning if a similar factorial structure could be found in comparison to previous studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020), a parallel analysis was done to check whether a similar factor structure as Balaji and Borsci (2019) using the Psych package (Revelle, 2019). Factors were extracted based on their eigenvalue, which should be above one. A parallel analysis was performed on a chatbot x item dataset and participant x item dataset to inform decision making regarding the number of factors. A principal component analysis was performed using the Psych package (Revelle, 2019). A Promax rotation was used, since this is an oblique rotation, which means that the factors can correlate and are not independent (Field, 2013).

For the second research question, the goal was to see if the 42-item BotScale could be reduced to a more manageable number of items with acceptable reliability. First, items with a factor loading lower than 0.3 or if the item was cross-loading on multiple items, were deleted. According to Field (2013), 0.3 could be sufficient but also depends on the sample size. Subsequently, all the items were evaluated to see which items should be removed. This was done by checking the density plot of each item and the mean of every item. Furthermore, reliability was measured using the Psych package (Revelle, 2019). If these steps did not reduce the dataset enough, similarly to Silderhuis and Borsci (2020), per chatbot feature the item with the highest factor loading would remain and the rest would be deleted. Lastly, a reliability analysis was performed on the shortened BotScale, and it will be compared to the shortened scales of Borsci et al. (2015) and Silderhuis and Borsci (2020).

The third goal was to explore the relationship between the Dutch version correlates and the English version of the BotScale. To achieve this a non-parametric correlation analysis (Kendall's Tau)

was performed accounting for the usage of Likert scales (Field, 2013). Additionally, group means were compared using Bayesian regression model stan_glm from the R_stanarm package (Gabry, Goodrich, Ali, & Brilleman, 2020). This method is chosen since using the comparison of groups (CGM) method allows two compare two groups, where one is seen as the default group to compare the other group to (Schmettow, 2020). In this case, the default was the English version of the questionnaire. Thereafter, ggplot2 was used to plot the data (Wickham, 2016).

Moreover, a correlation analysis (Kendall's Tau) was performed to answer the last research question to explore the relationship between the BotScale and the SUISQ-R.

3. Results

The following section will be divided into analyses on the BotScale, analyses on the SUISQ-R and ending with comparing the BotScale and SUISQ-R. In Appendix G, the R script can be found.

3.1 The BotScale

3.1.1 Parallel analysis

The parallel analysis was done on the whole dataset to observe if a similar factor structures to one identified in previous studies could be replicated (Balaji & Borsci, 2019). The scree plot, showing the chatbot x item analysis (Figure 1) and the scree plot showing the analysis of participant x item (Figure 2) suggested between three and five factors, which is in line with the four factor loading identified by Balaji and Borsci (2019).

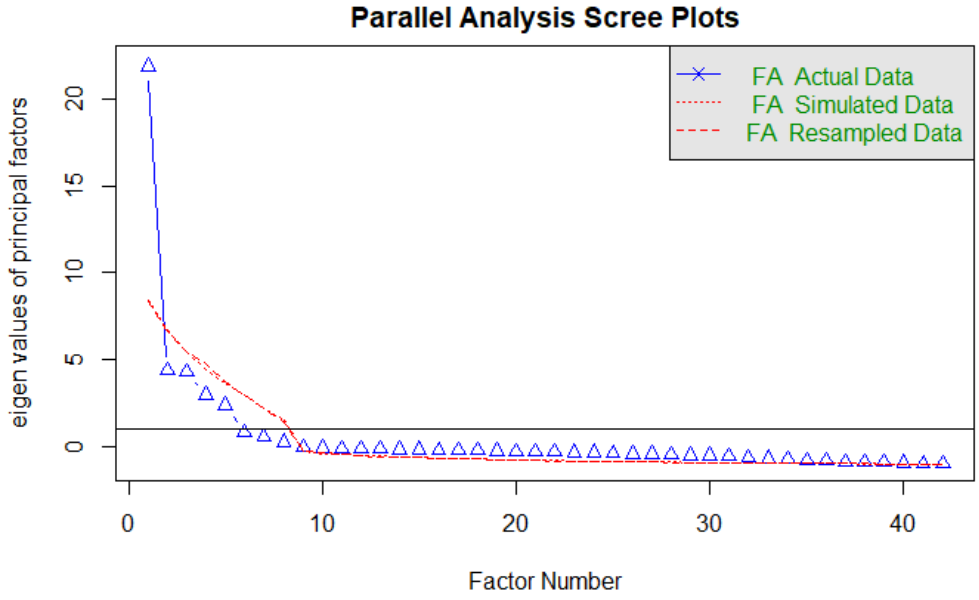

Figure 1 Scree plot resulting from a parallel analysis for the chatbot x item dataset, including both the Dutch and English version.

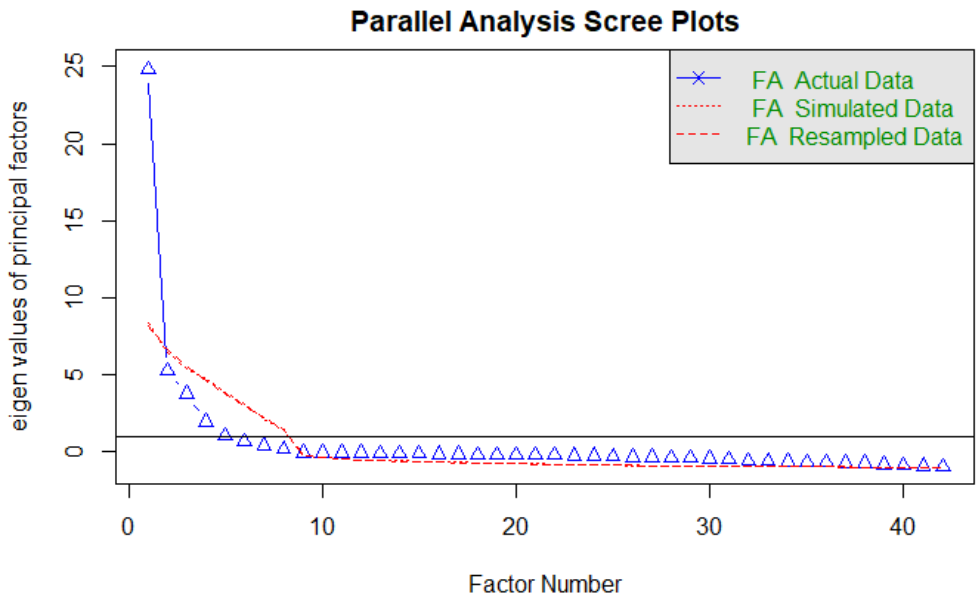

Figure 2 Scree plot from parallel analysis performed on the participant x item dataset for the BotScale, on both the English and Dutch version.

3.1.2. Principal component analysis on a four-factor structure

The results from the Principal Component Analysis (PCA) performed on the whole dataset can be found in Table 1. There were several factors which loaded almost equally on two factors and there were no other items which had a loading lower than 0.5. This factor structure explains 85% of the variance. Furthermore, the eigenvalues of the factors are all above five, with RC1 having the highest eigenvalue, namely 19.17.

Table 1

The loadings and eigenvalues of the Principal Component Analysis (PCA) on the means of every item per chatbot, on the full dataset. RC1 is Communication quality (Factor 2), RC2 is Conversation start and privacy (F1). RC3 is Graceful breakdown (F2), and RC4 for Perceived speed(F4).

Item	RC1	RC2	RC3	RC4
7	0.697			
9	0.594			
10	0.950			
11	1.015			
12	1.040			
14	0.920			
15	1.041			
16	0.698			
19	0.936			
22	0.979			
23	0.840			
24	0.875			
25	1.061			
26	0.826			
27	0.784			
28	0.963			
29	0.840			
30	0.975			
31	0.856			
34	0.764			
35	0.639			
36	*0.562*		*0.534*	
37	0.622			
38	0.690			
39	0.676			
1		0.907		
2		0.813		
3		0.732		
4		0.832		
5		0.954		
6		0.979		
17	*-0.343*	*-0.444*		*0.307*
18		*-0.444*	*0.346*	
20		0.688		
13			0.821	
32			0.917	
33			0.890	
8	*0.417*			*0.485*
21	*0.647*			*0.694*
40				0.995
41				1.008
42				0.999
Eigenvalues	19.170	5.865	5.648	5.373
Variance	0.456	0.140	0.134	0.128

Preceding the current study, Balaji and Borsci (2019) and Silderhuis and Borsci (2020) have also performed an exploratory factor analysis on the BotScale. A comparison of the structure can be seen in Table 2. If items with a loading of 0.5 and below are left out, it can be seen that the first factor is almost the same for all three studies. The current study and Silderhuis and Borsci (2020) have Q10 and Q11 both in factor two instead of one. Additionally, the current study has Q20 in factor one, while the other studies placed it in factor three. For the second factor, the current study does not include Q18, Q33, while the others do. The current study and Silderhuis and Borsci (2020) both do not include Q32 and Q36 (or with uncertainties), and only Silderhuis and Borsci (2020) does not include Q38. For factor three, the current study and Balaji and Borsci (2019) both have Q13, while Balaji and Borsci (2019) and Silderhuis and Borsci (2020) both have Q19, Q20 and Q21. For factor four, all studies include Q40, Q41, Q42, while the current study includes Q21 and Silderhuis and Borsci (2020) includes Q36.

Table 2

Comparison of preceding studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020) and the current study on the factor structure and the distribution of the items. In bold are items from the current study which are in a different category from both the previous studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020).

	Balaji and Borsci (2019)	Silderhuis and Borsci (2020)	Current study
F1	Q1, Q2, Q3, Q4, Q5, Q6, Q10, Q11	Q1, Q2, Q3, Q4, Q5, Q6	Q1, Q2, Q3, Q4, Q5, Q6, **Q17*, Q18*, Q20**
F2	Q7, Q8, Q9, Q12, Q14, Q15, Q16, Q17, Q18, Q22, Q23, Q24, Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q32, Q33, Q34, Q35, Q36, Q37, Q38, Q39	Q7, Q8, Q9*, Q10, Q11, Q12, Q13, Q14, Q15, Q16, Q18, Q22, Q23, Q24, Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q33*, Q34, Q35, Q37, Q39	Q7, Q9, Q10, Q11, Q12, Q14, Q15, Q16, **Q19**, Q22, Q, 23, Q24, Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q34, Q35, Q36**, Q37, Q38, Q39
F3	Q13, Q19, Q20, Q21	Q19, Q20, Q21, Q32*, Q38*	Q13, Q, 32, **Q33**
F4	Q40, Q41, Q42	Q36, Q40, Q41, Q42	**Q8*, Q21****,** Q40, Q41, Q42

* Items with a factor loading below 0.5

**Also loads above 0.5 on another factor

3.1.3 Item evaluation BotScale

To explore the possibility to reduce the number of items, with minimal effects on the reliability of the scale, items were removed following these exclusion criteria: i) low factor loadings, ii) items with little variance by looking at means above four or below two, iii) items with a spread which is only distributed across two scale points. Moreover, the reliability of the scale after each item that was dropped was estimated as well as item-total correlations, to decide to drop or retain an item. However, something to keep in mind is that each chatbot feature, which were identified by Tariverdiyeva and Borsci (2019) and refined by (Balaji & Borsci, 2019), to still be represented by at least one item.

Starting with looking at the factor loadings, there are only a few items with loadings below 0.5, which are item Q8, Q17 and Q18. However, these items also load onto another factor, which is another reason to exclude these items. Item Q21 and Q36 have loading above 0.5 but are also loading onto multiple factors with close factor loadings. Since the gaps are both below 0.1, these items were also deleted.

Subsequently, the means score of each item and the spread was looked at. An item with a very high or low mean could indicate that this item does not explain much variance between chatbots. However, it can also be that only a few scale points were used to answer an item by each participant, again indicating that this item does not differentiate a lot between chatbots. There were no items with a mean lower than two, however there were items with a mean higher than four: Q1, Q2, Q3, Q4, Q39, Q40, Q41, Q42. Additionally, Q38 had a mean of 3.99, which is also doubtful. At the same time, the spread of the items was also considered. The distribution of the answers per item can be found in Appendix G. There were a few items of which the scores were only distributed over two scale points or less. These items were: Q1, Q2, Q3, Q19, Q33, Q38. And items Q40, Q41, Q42 do have some variance but have a few steep peaks instead of one wider peak like other items. Items Q1, Q2, Q3 cannot all be deleted, since then the feature “Ease of starting a conversation” will not be represented. Since Q1 has the highest factor loading, this item will not be deleted. Similarly, this counts for Q40, Q41 and Q42, which all are tied to the feature “Perceived speed”. Here, Q41 will not be deleted since it has the highest factor loading and lowest mean out of the three.

Lastly, reliability if an item is dropped will be considered for the selection of items. The items

which were already reviewed as “bad”, are already taken out of the dataset. The overall reliability of the scale is high with a Cronbach $\alpha = 0.977$, with F1 $\alpha = 0.889$, F2 $\alpha = 0.990$, F3 $\alpha = 0.723$ and F4 has only one item left. In this step, there were no items that could be deleted when looked at the reliability if an item was dropped. If this was the only criteria, item Q1, Q20, Q13 and Q41 would have been dropped, however then the corresponding feature would not be represented anymore. Item-total correlations were also looked at, however no additional items could be deleted based on them.

In total, 14 items were deleted in these steps and thus 28 items remain after the above taken steps. However, this scale could still be reduced to 14 items in total, if one item per chatbot feature would be chosen. Silderhuis and Borsci (2020) did this by looking at the items of the feature and selecting the item with the highest factor loading. In the current study, this method was also used, and the resulting item list can be found in Table 3. Additionally, a comparison to the shortened scales of Balaji and Borsci (2019) and Silderhuis and Borsci (2020) can be seen here. There were four features where all studies chose the same factor. For four features of the current study there was one other study agreeing with the item. For three features the other studies were agreeing with each other. Meaning that for three features, all the studies disagreed on the item. The 14-item shortened version of the scale had a Cronbach’s $\alpha = 0.93$.

Table 3

Showing the shortened version of the BotScale and comparing it to the two previous theses who also made a shortened version. In bold are marked the items which were placed in another factor for the previous studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020).

Factor	Feature	Present study	Balaji and Borsci (2019)	Silderhuis and Borsci (2020)
F1 Conversation start and privacy	Ease of starting a conversation	Q1	Q1, Q2	Q2
	Accessibility	Q6	Q4, Q5	Q5
	Perceived privacy	Q20	**Q21 (F3)**	**Q19 (F3)**
F2 Communication quality	Expectation setting	Q7	Q7	Q7
	Communication effort	Q12	**Q10, Q11 (F1)**	Q10
	Ability to maintain themed discussion	Q15	Q15	Q15
	Reference to service	Q16	Q18	Q16
	Recognition and facilitation of user's goal and intent	Q22	Q24	Q24
	Relevance	Q25	Q25	Q27
	Maxim of quantity	Q28	Q30	Q29
	Understandability	Q34	Q34	Q34
	Perceived credibility	Q37	Q37	Q37
F3 Graceful breakdown	Graceful breakdown	Q32	**Q33 (F2)**	**Q31 (F2)**
F4 Perceived speed	Perceived speed	Q41	Q41	Q42

3.1.4 Comparing Dutch and English version of the BotScale

The 14-item version of the Dutch and English chatbot scale were assessed using a Compare Group Means Analysis (see Figure 3). Kendall's rank correlation showed a positive correlation with $p = 0.002$ and $\tau = 0.82$. Furthermore, the English version has a Cronbach's alpha of $\alpha = 0.95$ and the Dutch version a Cronbach's alpha of $\alpha = 0.89$.

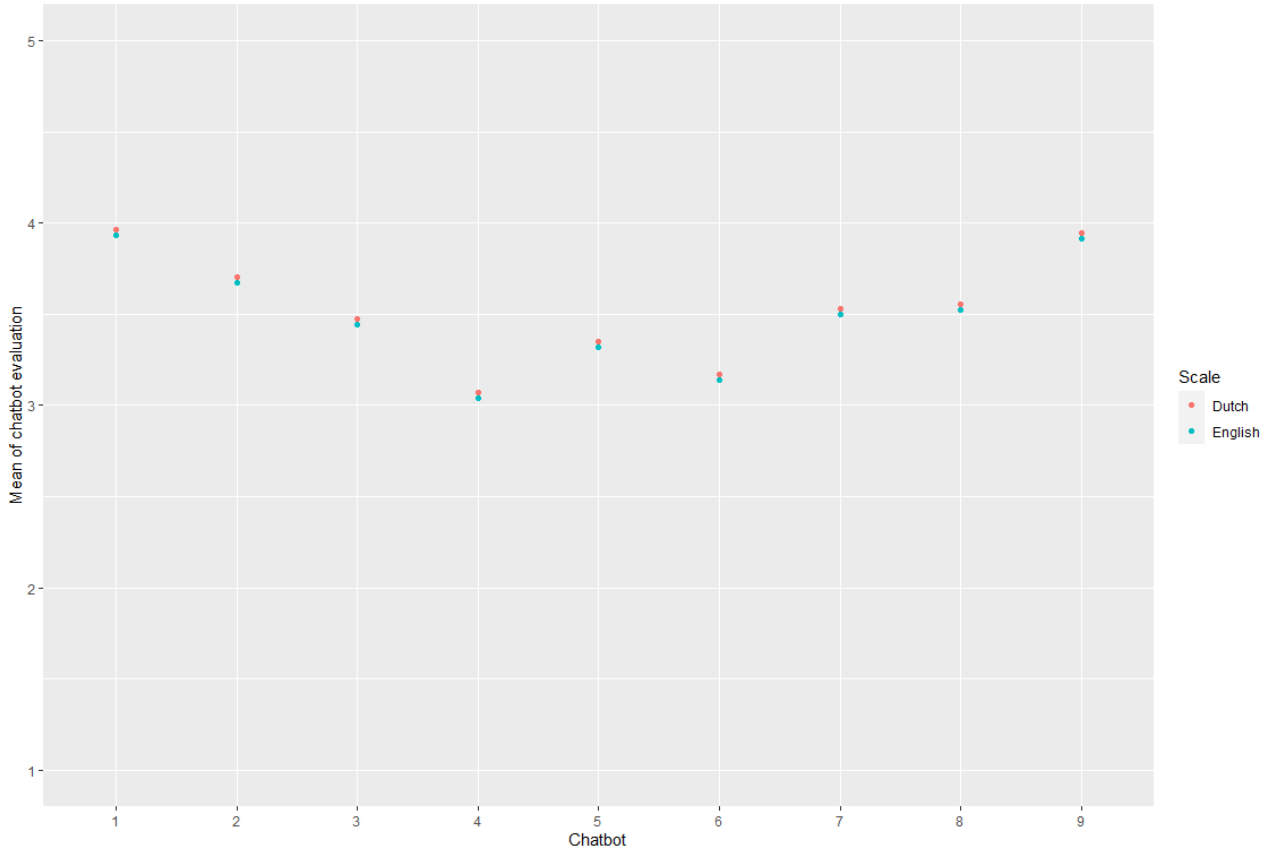

Figure 3 Plot showing the results of a Compare Group Means analysis on the average score of the satisfaction from the BotScale.

3.2 SUISQ-R

3.2.1 SUISQ-R translation

The results from the reliability analysis for each factor of the English and Dutch version of the SUISQ-R can be seen in Table 4. Since the questions about the voice of the system were left out, "Speech Characteristics" is left blank and not considered in the analysis. The results from the original study by Lewis and Hardzinski (2015) about the SUISQ-R can also be seen for comparison. The overall reliability for the English and the Dutch version were more than acceptable, respectively Cronbach's $\alpha = 0.89$ and Cronbach's $\alpha = 0.84$.

Table 4

Results of the reliability analysis using Cronbach's alpha on the English and Dutch version of the SUISQ-R, without the voice items. Additionally, a comparison is given with the results from the study of Lewis and Hardzinski (2015) where the SUISQ was reduced to the SUISQ-R and analysed. The current study administered the SUISQ-R using a five-point Likert scale, while the original study used a seven-point Likert scale.

	English	**Dutch**	**(Lewis & Hardzinski, 2015)**
Complete	0.92	0.83	0.88
User Goal Orientation	0.95	0.91	0.91
Customer Service Behaviour	0.67	0.47	0.88
Speech Characteristics	- not tested	- not tested	0.80
Verbosity	0.52	0.41	0.67

3.2.2 Relationship between the BotScale and SUISQ-R

A Kendall rank correlation analysis between the BotScale and the SUISQ-R without the voice items showed a positive correlation with $p = 0.01$ and $\tau = 0.78$. A graphical representation of the means of both scales can be seen in Figure 4. Furthermore, Kendall rank correlations between the factors of the BotScale and the SUISQ-R and its subscales can be seen in Table 5.

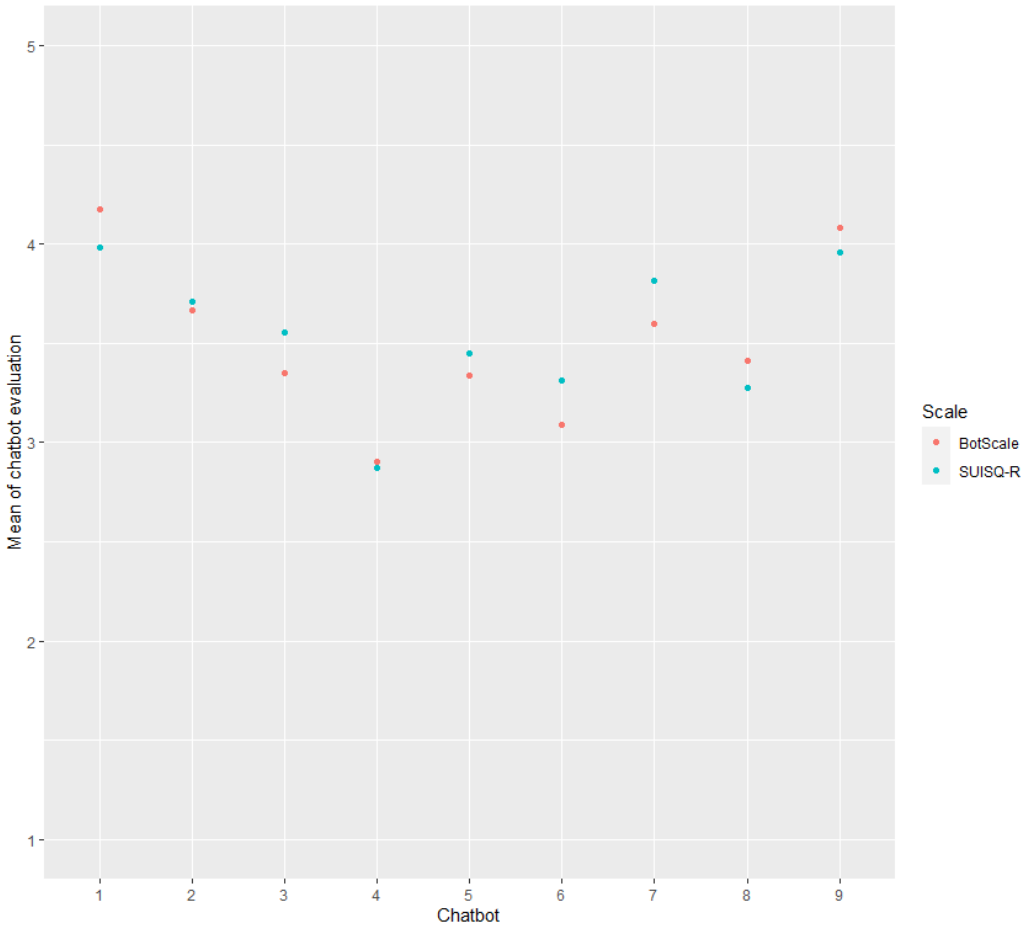

Figure 4 Showing the difference in means resulting from the evaluation after interacting with each chatbot for the BotScale and the SUISQ-R (Lewis & Hardzinski, 2015).

Table 5

*Results of a Kendall rank correlation analysis between the BotScale and SUISQ-R (Lewis & Hardzinski, 2015) and their subscales. The subscales of the SUISQ-R are: User Goal orientation (UGO), Customer Service Behaviour CSB) and Verbosity (V). Speech Characteristics is another one of their subscales but was left out since these questions are specifically about a voice (Lewis & Hardzinski, 2015). * p < 0.05, **p<0.001.*

	SUISQ-R	**UGO**	**CSB**	**V**
BotScale	0.778*	0.833**	0.611*	0.500
F1	0.222	0.056	0.167	0.056
F2	0.778*	0.833**	0.611*	0.500
F3	0.261	0.261	0.203	0.145
F4	-0.028	0.141	-0.084	-0.197

4. Discussion

This study is aimed to replicate the study of Balaji and Borsci (2019), to see if a similar factor structure could be identified. Additionally, in line with previous studies on this new tool, the BotScale was shortened and compared to the results of Balaji and Borsci (2019) and Silderhuis and Borsci (2020). An existing scale for user satisfaction in voice interfaces, the SUISQ-R (Lewis & Hardzinski, 2015), was also included in the survey for participants. Since this scale was currently only available in Dutch, it was first translated and then a reliability analysis was done. To perform an external validation of the proposed BotScale, a correlation analysis was done between the BotScale and the SUISQ-R.

4.1 Factorial structure

The first research question was: "Is the factorial structure of the questionnaire in line with previous studies?" From the principal component analysis, it could be seen that a structure with four factors can explain 85% of the variance. However, according to Cangelosi and Goriely (2007) there is no standard rule for the amount of variance needed which we should follow, since the amount of variance needed can change per study and subject. The cumulative variance is higher than the one proposed by Silderhuis and Borsci (2020), where 57,6 % could be explained with a similar four factor structure. When comparing the distribution of the items, with the distributions of the previous studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020), an almost similar structure can be found by confirming the organisation of the scale in four factors. Therefore, it seems a four factor structure seems like a good fit, since 85% of the variance is explained and it is comparable to previous work.

4.2 Shortened BotScale

Secondly: "Can the BotScale be shortened, while keeping high reliability and including items about every chatbot feature?". To achieve this, items were evaluated based on their factor loadings and cross-loadings. Subsequently, items were evaluated on their means and distribution and lastly on their reliability and item-total correlation. This resulted in a scale which had 28 items, however the number of items could be further reduced. Therefore, the item with the highest factor loading per chatbot feature was retained, similar to how Silderhuis and Borsci (2020) reduced their number of items. This

resulted in having one item per chatbot feature, thus having 14 items (See Appendix H). Overall, the shortened scale still had a strong reliability with a Cronbach's $\alpha = 0.93$. In general, the cut-off for Cronbach's alpha is 0.70. So, the 14-item BotScale seems to have a good reliability. However, a Cronbach's alpha above 0.90 can sometimes indicate redundancy within the scale (Lavrakas, 2008). Therefore, it may be important to see if the scale can be reduced even further, however the scale currently only had one item per chatbot feature. It could be that some of the chatbot features are interpreted similarly by the user or that some chatbot features are correlated to each other.

4.3 Translation BotScale

The third research question was "Does the Dutch translation of the BotScale correlate with the original version?". The results showed a significant positive correlation when comparing the shortened Dutch version to the shortened English version. A correlation coefficient is seen as moderate when ranging between 0.4-0.69 and strong from 0.7-0.9 (Akoglu, 2018). Thus, a strong correlation was found for the translated and original scale. Additionally, when looking at the Bayesian regression model which was used to compare the group means, only a small difference can be seen for the mean satisfaction score for each chatbot. This can indicate that the Dutch translation captures the essence of the original wording of the scale and similarily measures user satisfaction.

4.4 Comparing the SUISQ-R and BotScale

The last research question was: "Do the BotScale and SUISQ-R show a moderate to strong correlation when comparing the ratings of the chatbots?". Reliability of the SUISQ-R was measured using Cronbach's alpha of both the English and Dutch version and subsequently they were compared to the original results of Lewis and Hardzinski (2015). The reliability of the Dutch and English scales was lower for two of the subscales when both were compared to Lewis and Hardzinski (2015), with the biggest difference being in "Customer Service Behaviour" of around 0.40 between the original and Dutch version.

Then to answer the research question, a correlational analysis was done. When doing the correlational analysis, a significant positive moderate correlation was found. A moderate correlation instead of a strong correlation could be expected since a scale for voice interfaces is compared to a

scale for chatbots. In some ways they are similar since they both interact with humans. However, the way they interact is different, so this could be why there is only a moderate correlation and not a strong correlation. Furthermore, when comparing the factors of the BotScale independently with the SUISQ-R and its factors, several things can be noticed. Factor 2 has the strongest correlation to the SUISQ-R and is also similar to the correlation between the SUISQ-R and the overall BotScale. However, Factor 2 contains the most items, namely nine, while Factor 3 contains three items and the other two factors both only contain one item. The other three factors do not have a moderate or strong correlation to the SUISQ-R. This could be because e.g. Factor 1 has items asking about accessibility or privacy, Factor 3 about a graceful breakdown, and these features are not included in the SUISQ-R. Thus it seems the SUISQ-R does not cover some of the features which seem important to evaluate chatbots, which were found by systematic literature reviews and focus groups in preceding work (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019).

In previous studies, the UMUX-lite (Lewis et al., 2013) was compared to the BotScale (Silderhuis & Borsci, 2020; Tariverdiyeva & Borsci, 2019). Silderhuis and Borsci (2020) did find a strong correlation when comparing the BotScale with the UMUX-lite, however when comparing the individual subscales of the BotScale with the UMUX-lite three out of four subscales had a weak correlation. This could indicate that the UMUX-lite does not contain all important aspects to evaluate chatbots, which is also what Tariverdiyeva and Borsci (2019) concluded in their study.

4.5 Limitations of the present study

A first limitation of the current study is that the sample size only included people aged from around 18-30. According to a study from Friemel (2014), of seniors (> 65 years old) in Switzerland only 35,9% had used the internet in the past half-year. They also researched the reasons why some seniors did not use the internet. There were many different reasons, but the most agreed upon reasons were that it was difficult to learn but would also cost a lot of effort. They were concerned about their safety or afraid to encounter problems. Most reasons mentioned were psychological, but 30% also mentioned their degraded vision or hearing (Friemel, 2014). Thus, it is important to also take the opinions of this age group into account when developing a chatbot scale. It might even be that they also find other

things important in a chatbot design, e.g. readability. However, Silderhuis and Borsci (2020) also did a replication study of Balaji and Borsci (2019), but their sample included two participant groups aged 25-35 and 55-70. They found that comparing the results of the two groups, the BotScale showed only minimal differences regarding two featurs, concluding that one version is sufficiently robust to accommodate participants of different age for both age groups.

Aditionally, the sample of the current study did not have many people participating who had experience with chatbots. Demographic questions were asked about how familiar participants were with chatbots. People were on average moderately familiar and had probably or definetely seen a a chatbot before. However, the average participant never or rarely used a chatbot. Thus this study does not include expert opinions on the scale.

Due to the COVID19 pandemic, the study had to be executed online which brought some limitations. People had to use their own computers, so the experience with chatbot was affected by different techical setting, for instace some participants cookie settings prevented access to some of the chatbots. And participants were forced to interact while in incoginto mode to then be able to interact with the chatbot. Since everyone was working at home, sometimes there were small distractions since other people accidentally entered the room, there where noise from outside etc. So, it was harder to try to create a controlled environment. There were also a few cases where the camera of the participant stopped working. This should not matter much, since the researcher could still hear the participant at all time and could track their progress in the survey.

The last limitation had to do with the fact that participants only had to perform one task per chatbot and then answer a very long set of repetitive questions after each task resulting in a long procedure that could be considered quite demanding for participants. Additionally, one task per chatbot may not be enough to get to evaluate the chatbots on all aspects.

4.6 Future research

For future research, a suggestion would be to include participants who have more experience with chatbots and use them regularly or even develop them. It would be interesting to see if they would evaluate chatbots the same and a similar factor structure will be reached. This is also important, since

experts have probably seen multiple chatbots, use them regilarily or design them, which could give them a different perspective on what makes a chatbot with high usability.

The current study and two previous studies (Balaji & Borsci, 2019; Silderhuis & Borsci, 2020) have reduced the BotScale, which yielded fairly similar results. Therefore, further testing could be done with the reduced version, which could also focus more on bigger or more diverse samples or more elaborate tasks. In the current study, participants only had one task to get to know the chatbot and to evaluate them on. In some cases this meant an interaction only took one minute. It could be that participants would need some more interaction to get a better impression of the chatbot. Therefore, a suggestion is to have tasks for each chatbot, which make sure that every chatbot feature will be tested. Additionally, to minimise the strain on participants, the reduced 14-item scale should be used in combination with more or longer tasks.

5. Conclusion

This study has shown that a similar four-factor structure, item distribution and shortened 14-item chatbot usability scale to previous works can be achieved while maintaining high reliability of $\alpha = 0.93$. This scale can be used to evaluate a chatbot on user satisfaction, which is important since more and more websites may start using and developing chatbots. Additionally, the Dutch translation of the scale had a strong positive correlation when it was similarly reduced as the English version. The SUISQ-R was also translated to Dutch, which made it possible to do a correlational analysis between the SUISQ-R and the BotScale. A moderate positive correlation was found between the two scales, which could be expected since the SUISQ-R is developed for voice interfaces. Overall, this study showed that the BotScale can be shortened, is still reliable, has a Dutch translation and shows a moderate correlation to an existing voice interface scale. However, it should be kept in mind that chatbot experts were not represented in the sample and that the tasks could be improved so that participants get a better impression of the chatbots they are evaluating.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 100006. doi:https://doi.org/10.1016/j.mlwa.2020.100006

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91-93. doi:https://doi.org/10.1016/j.tjem.2018.08.001

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183-189. doi:https://doi.org/10.1016/j.chb.2018.03.051

Balaji, D., & Borsci, S. (2019). *Assessing user satisfaction with information chatbots: A preliminary investigation.* (Master thesis). University of Twente, Enschede, Netherlands.

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., . . . Moreira, C. (2020). Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review, 36*, 100239. doi:https://doi.org/10.1016/j.cosrev.2020.100239

Ben Mimoun, M. S., Poncin, I., & Garnier, M. (2017). Animated conversational agents and e-consumer productivity: The roles of agents and individual characteristics. *Information & Management, 54*(5), 545-559. doi:https://doi.org/10.1016/j.im.2016.11.008

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human–Computer Interaction, 31*(8), 484-495. doi:10.1080/10447318.2015.1064648

Botanalytics. (2018). How chatbot performance metrics differ by industry. Retrieved from https://chatbotslife.com/how-chatbot-performance-metrics-differ-by-industry-b380d4bd7f6b

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. . In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evolution in industry* (pp. 189-194). London, UK: Taylor & Francis.

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology direct, 2*, 2. doi:10.1186/1745-6150-2-2

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research, 117*, 587-595. doi:https://doi.org/10.1016/j.jbusres.2018.10.004

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4 ed.). London, England: SAGE Publications Ltd.

Finstad, K. (2010). The Usability Metric for User Experience. *Interacting with Computers, 22*(5), 323-327. doi:https://doi.org/10.1016/j.intcom.2010.04.004

Friemel, T. N. (2014). The digital divide has grown old: Determinants of a digital divide among seniors. *New Media & Society, 18*(2), 313-331. doi:10.1177/1461444814538648

Gabry, J., Goodrich, B., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan (Version 2.21.1). Retrieved from https://mc-stan.org/rstanarm/

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304-316. doi:https://doi.org/10.1016/j.chb.2019.01.020

Google Meet. (n.d.). Retrieved from https://apps.google.com/meet/

Hone, K., & Graham, R. (2000). Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering, 6*. doi:10.1017/S1351324900002497

Interaction Design Foundation. (n.d.). Voice User Interfaces. Retrieved from https://www.interaction-design.org/literature/topics/voice-user-interfaces

Io, H. N., & Lee, C. B. (2017, 10-13 Dec. 2017). *Chatbots and conversational agents: A bibliometric analysis.* Paper presented at the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).

Jenkins, M.-C., Churchill, R., Cox, S., & Smith, D. (2007). Analysis of User Interaction with Service Oriented Chatbot Systems. In J. A. Jacko (Ed.), *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. HCI 2007. Lecture notes in Computer Science* (Vol. 4552, pp. 76-83). doi:https://doi.org/10.1007/978-3-540-73110-8_9

Lavrakas, P. J. (2008). Cronbach's alpha. In *Encyclopedia of survey research methods* (Vol. 1-0). Thousand Oaks, CA: Sage Publications, Inc.

Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies, 103*, 95-105. doi:https://doi.org/10.1016/j.ijhcs.2017.02.005

Lewis, J. (2018). Investigating MOS-X Ratings of Synthetic and Human Voices. *Voice Interaction Design, 2*(2), 1-22. Retrieved from https://www.researchgate.net/publication/330224885_Investigating_MOS-X_Ratings_of_Synthetic_and_Human_Voices

Lewis, J., & Hardzinski, M. (2015). Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology, 18*. doi:10.1007/s10772-015-9289-1

Lewis, J., & Sauro, J. (2020). Three questionnaires for measuring voice interaction experiences [Blog post]. Retrieved from https://measuringu.com/voice-interaction/

Lewis, J., Utesch, B. S., & Maher, D. E. (2013). *UMUX-LITE: when there's no time for the SUS.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France. https://doi.org/10.1145/2470654.2481287

Polkosky, M. D. (2005). *Toward a Social-Cognitive Psychology of Speech Technology: Affective Reponses to Speech-Based e-Service.* (Dissertation). University of South Florida, Retrieved from https://scholarcommons.usf.edu/etd/819/

Polkosky, M. D., & Lewis, J. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology, 6*, 161-182. doi:10.1023/A:1022390615396

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., & Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons, 62*(6), 785-797. doi:https://doi.org/10.1016/j.bushor.2019.08.005

Qiu, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems, 25*(4), 145-182. doi:10.2753/MIS0742-1222250405

Qualtrics. (n.d.). Retrieved from https://www.qualtrics.com/

R Core Team. (2020). R: A language and evironment for statistical computing (Version 4.0.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Revelle, W. (2019). psych: Procedures for personality and psychological research. (Version 2.0.7). Retrieved from https://www.personality-project.org/

Sajjadi, P., Hoffmann, L., Cimiano, P., & Kopp, S. (2019). A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users. *Entertainment Computing, 32*, 100313. doi:https://doi.org/10.1016/j.entcom.2019.100313

Schmettow, M. (2020). *New Statistics for design researchers*. Retrieved from https://schmettow.github.io/New_Stats/index.html

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F. (2018). The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems, 114*, 94-102. doi:https://doi.org/10.1016/j.dss.2018.08.011

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research, 115*, 14-24. doi:https://doi.org/10.1016/j.jbusres.2020.04.030

Silderhuis, I., & Borsci, S. (2020). *Validity and reliability of the user satisfaction with Information Chatbots Scale (USIC).* (Master Thesis). University of Twente, Enschede, The Netherlands.

Tariverdiyeva, G., & Borsci, S. (2019). *Chatbots' perceived usability in information retrieval tasks: An exploratory analysis.* (Master thesis). University of Twente, Enschede, The Netherlands.

van Os, R., Hachmang, D., Akpinar, M., Kreuning, A., & Derksen, M. (2018). *Stand van webcare 2018*. Retrieved from https://www.upstream.nl/wp-content/uploads/2018/09/20180918-Onderzoek-Stand-van-Webcare-2018.pdf

Wang, W., & Benbasat, I. (2008). Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce. *Journal of Management Information Systems, 24*(4), 249–273. doi:10.2753/mis0742-1222240410

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113-117. doi:https://doi.org/10.1016/j.jesp.2014.01.005

Wickham, W. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Zamora, J. (2017). *I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations*. Paper presented at the Proceedings of the 5th International Conference on Human Agent Interaction, Bielefeld, Germany. https://doi.org/10.1145/3125739.3125766

Appendices

Appendix A: Three voice interface scales (MOS-X, SASSI, SUISQ)

The shortened versions of the scales are also included. The shortened version of the MOS-X had rewritten items and are thus included below the MOS-X. The SUISQ had two shorter versions, a 14 item version **SUISQ-R(bold)** and a 9 item version ***SUISQ-MR (bold and cursive).***

MOS- X (Polkosky & Lewis, 2003)	**SASSI (Hone & Graham, 2000)**	**SUISQ (Lewis & Hardzinski, 2015; Polkosky, 2005)**
Please rate the degree of effort you had to make to understand the message.	The system is accurate.	**The system made me feel like I was in control.**
Were single words hard to understand?	The system is unreliable.	***The messages were repetitive.***
Were the speech sounds clearly distinguishable?	The interaction with the system is unpredictable.	The system gave me a good feeling about being a customer of this business.
Was the articulation of speech sounds precise?	The system didn't always do what I wanted.	The system used terms I am familiar with.
Was the voice you heard pleasant to listen to?	The system didn't always do what I expected.	**I could find what I needed without any difficulty.**
Did the voice sound natural?	The system is dependable.	***The system used everyday words.***
To what extent did this voice sound like a human?	The system makes few errors.	The system was organized and logical.
Did the voice sound harsh, raspy or restrained?	The interaction with the system is consistent.	The system gave me more details than I needed.
Did emphasis of important words occur?	The interaction with the system is efficient.	The system spoke at a pace that was easy to follow.
Did the rhythm of the speech sound natural?	The system is useful.	The system would help me be productive.
Did the intonation pattern of sentences sound smooth and natural?	The system is pleasant.	***The system seemed polite.***
Did the voice appear to be trustworthy?	The system is friendly.	I could trust this system to work correctly.
Did the voice suggest a confident speaker?	I was able to recover easily from errors.	***I would be likely to use this system again.***
Did the voice seem enthusiastic?	I enjoyed using the system.	The system's voice was pleasant.
Was the voice persuasive?	It is clear how to speak to the system.	***The system was too talkative.***
	It is easy to learn to use the system.	The system's voice sounded like people I hear on the radio or television.
	I would use this system.	***I felt confident using this system.***
MOS-X2		
Please rate the extent to which it was easy or difficult to understand what the voice was saying.	I felt in control of the interaction with the system.	**The system's voice sounded like a regular person.**

How natural (pleasantly, human-like) was the sound of the voice? To what extent were the elements of timing, pitch and emphasis appropriate for the message? To what extent was the tone of the voice socially and emotionally appropriate for the messages?	I felt confident using the system. I felt tense using the system. I felt calm using the system. A high level of concentration is required when using the system. The system is easy to use. The interaction with the system is repetitive. The interaction with the system is boring. The interaction with the system is irritating. The interaction with the system is frustrating. The system is too inflexible. I sometimes wondered if I was using the right word. I always knew what to say to the system. I was not always sure what the system was doing. It is easy to lose track of where you are in an interaction with the system. The interaction with the system is fast. The system responds too slowly.	The quality of this system made me want to remain a customer of this business. ***The system's voice sounded natural.*** The system seemed courteous. ***I felt like I had to wait too long for the system to stop talking so I could respond.*** **The system seemed friendly.** ***The system's voice sounded enthusiastic and full of energy.*** **The system seemed professional in its speaking style.**

Appendix B: Informed consent

English version

Dear participant,
Thank you for participating in this research. I would like to tell you a few things before we get started to inform you properly. Firstly, remember that your participation is voluntarily, which also means that you can stop at any time without giving any reasons without there being negative consequences.

Purpose of the research
This research aims to create a scale, with which chatbots can be evaluated. A chatbot is a program with which you can chat trough text and it will give you answers based on what you say, for example used in customer service. Additionally, we have made a Dutch version of this chatbot scale and the voice interface scale, so we aim to compare it to the English version to see if it still conveys the same message.

Study content
You will receive some tasks from me, interact with a few chatbots and after this interaction with each chatbot you will have to fill out a scale to evaluate the chatbot and an additional existing scale for voice interfaces to compare our scale with. A voice interface can also interact with humans but does so by using a human like voice. The study will take around an hour to 75 minutes, and there are no risks attached to your participation.

Data acquisition
In the end, we hope to use this data to see which items in the scale are important and help to evaluate a chatbot. Then we end with a tool which everyone can use to evaluate their chatbots. If you agree we would like to record your voice and the video meeting. Additionally, before tasks start we will ask some questions about your age, gender, nationality and previous experience with chatbots which we use to see for what kind of population we collect data. We will make sure that the data we collect of you will not be traceable back to you and we will not share any data with third parties. Only my supervisors will be able to see the data. It is possible that the data will be published, however data that would be able to identify you will be removed. The data will be stored in a secure data storage from the university, to which only my supervisor will have access.

Contact
If you ever have any questions after this session has ended you can email me: m.a.vandenbos@student.utwente.nl and my supervisor can be reached at s.borsci@utwente.nl. For questions about the ethical approval and your rights you can reach ethicscommittee-bms@utwente.nl . This study is approved by the ethical committee of the Behavioural, Management and Social Sciences (BMS) of the University of Twente.

	Yes	No
I have read and understood the study information dated [21/05/2020], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.		
I consent voluntarily to be a participant in this study and understand that I can refuse to answer question and I can withdraw from the study at any time, without having to give reason.		
I understand that taking part in the study involves answering questions about my demographics, performing tasks and interacting with chatbots online, filling out two scales about each of the five chatbots I have interacted with online and have an online call with the researcher that is being recorded.		
I understand that information I provide will be used for a master thesis and possibly for a publication.		
I understand that personal information collected about me that can identify me, such as [e.g. my name or where I live] will not be shared beyond the study team.		
I agree to be audio and video recorded		
I give permission for the filling out of the scales and demographics questionnaire that I provide to be archived in a safe data repository so it can be used for future learning.		

Dutch version

Beste deelnemer,
Bedankt dat u mee wilt doen met mijn onderzoek. Voor we beginnen zal ik u meer vertellen over het onderzoek en uw rechten. Uw deelname aan dit onderzoek is vrijwillig, dus u mag ook op elk moment stoppen zonder een reden te geven en daar zullen dan geen negatieve consequenties voor zijn.

Doel van het onderzoek
Dit onderzoek heeft als doel om een test te ontwikkelen zodat chatbots geëvalueerd kunnen worden. Chatbots zijn programma's waarmee je als het ware kan praten en dan reageren ze zonder dat daar een mens bij is betrokken. Ook willen we testen of deze Nederlandse versie correct vertaald is vanuit het Engels en of de voice interface scale correct vertaald is.

Inhoud van het onderzoek
Straks krijgt u van mij verschillende taken en zal u die uitvoeren door een chatbot te gebruiken. Als u hiermee klaar bent krijgt u de vragenlijst die is ontwikkeld om chatbots te evalueren en een vragenlijst voor voice interfaces, die we dan kunnen vergelijken met de chatbot vragenlijst. Een voice interface is ook een soort chatbot maar die communiceert via een stem i.p.v. via tekst. Het onderzoek zal ongeveer en uur en een kwartier duren en er zijn geen risico's verbonden aan meedoen aan het onderzoek.

Data verwerking
We willen de data van dit onderzoek gebruiken om te kijken welke items van de vragenlijst belangrijk zijn bij het evalueren van een chatbot en een scale te maken die gebruikt kan worden om standaard chatbots te evalueren. Als u akkoord gaat zouden we graag uw stem en beeld op willen nemen. Ook voor de taken beginnen vraag ik over uw leeftijd, nationaliteit, geslacht en ervaring met chatbots zodat we kunnen zien van wat voor soort populatie wij data verzamelen. We zorgen ervoor dat data die gebruikt wordt in mijn thesis of in een publicatie niet terug kan leiden naar u en data wordt niet gedeeld met derden. Mijn begeleiders kunnen de data als enigen ook inzien. Het is mogelijk dat de data gebruikt wordt in een publicatie, maar dan wordt ervoor gezorgd dat dit niet naar u terug kan leiden. Verder word de data veilig op geslagen in een opslag binnen de universiteit waar alleen mijn supervisor bij kan.

Contact
Als u na het aflopen van het onderzoek nog vragen heeft dan kunt u mij mailen op: m.a.vandenbos@student.utwente.nl en mijn supervisor is te bereiken op s.borsci@utwente.nl. Voor vragen over de ethische goedkeuring of uw rechten kunt u mailen naar ethicscommittee-bms@utwente.nl. Dit onderzoek is goedgekeurd door de ethische commissie van de Behavioural, Management and Social Sciences (BMS) afdeling van Universiteit Twente.

	Ja	Nee
Ik heb de hiervoor gegeven informatie gelezen en begrepen, of het is mij voorgelezen. Ik heb vragen kunnen stellen over het onderzoek en mijn vragen zijn naar tevredenheid beantwoord.		
Ik ga er vrijwillig mee akkoord om een deelnemer te zijn in dit onderzoek en begrijp dat ik het recht heb om het beantwoorden van vragen te weigeren en dat ik op elk moment kan stoppen met het deelnemen aan dit onderzoek zonder daar een reden voor te geven.		
Ik begrijp dat deelnemen in dit onderzoek inhoud dat ik vragen zal invullen over mijn demografische gegevenens, interactie zal hebben met online chatbots, twee vragenlijsten zal invullen over elk van de vijf chatbots waarmee ik interactie heb gehad en dat ik een videogesprek zal hebben met de onderzoeker die opgenomen word.		
Ik begrijp dat de informatie die ik aanlever gebruikt zal worden voor een master scriptie en mogelijk voor een publicatie.		
Ik begrijp dat mijn persoonlijke informatie welke mij zou kunnen identificeren, bijvoorbeeld mijn naam of waar ik woon, niet buiten het onderzoeksteam word gedeeld.		
Ik ga ermee akkoord dat mijn audio en video word opgenomen.		
Ik geef toestemming dat mijn antwoorden op de vragenlijsten en demografische vragen worden gearchiveerd in een veilige database zodat het gebruikt kan worden voor toekomstig onderzoek en leren.		

Appendix C: BotScale (Original/English)

BotScale from Balaji and Borsci (2019):

Could be answered on a five-point Likert scale ranging from: Strongly disagree (1) – Strongly agree

Respond to the next statements based on your experience with the chatbot:

1	It was clear how to start a conversation with the chatbot.
2	It was easy for me to understand how to start the interaction with the chatbot.
3	I find it easy to start a conversation with the chatbot.
4	The chatbot was easy to access.
5	The chatbot function was easily detectable.
6	It was easy to find the chatbot.
7	Communicating with the chatbot was clear.
8	I was immediately made aware of what information the chatbot can give me.
9	It is clear to me early on about what the chatbot can do.
10	I had to rephrase my input multiple times for the chatbot to be able to help me.
11	I had to pay special attention regarding my phrasing when communicating with the chatbot.
12	It was easy to tell the chatbot what I would like it to do.
13	The interaction with the chatbot felt like an ongoing conversation.
14	The chatbot was able to keep track of context.
15	The chatbot maintained relevant conversation.
16	The chatbot guided me to the relevant service.
17	The chatbot is using hyperlinks to guide me to my goal.
18	The chatbot was able to make references to the website or service when appropriate.
19	The interaction with the chatbot felt secure in terms of privacy.
20	I believe the chatbot informs me of any possible privacy issues.
21	I believe that this chatbot maintains my privacy.
22	I felt that my intentions were understood by the chatbot.
23	The chatbot was able to guide me to my goal.
24	I find that the chatbot understands what I want and helps me achieve my goal.
25	The chatbot gave relevant information during the whole conversation.
26	The chatbot is good at providing me with a helpful response at any point of the process.

27	The chatbot provided relevant information as and when I needed it.
28	The amount of received information was neither too much nor too less.
29	The chatbot gives me the appropriate amount of information.
30	The chatbot only gives me the information I need.
31	The chatbot could handle situations in which the line of conversation was not clear.
32	The chatbot explained gracefully when it could not help me.
33	When the chatbot encountered a problem, it responded appropriately.
34	I found the chatbot's responses clear.
35	The chatbot only states understandable answers.
36	The chatbot's responses were easy to understand.
37	I feel like the chatbot's responses were accurate.
38	I believe that the chatbot only states reliable information.
39	It appeared that the chatbot provided accurate and reliable information.
40	The time of the response was reasonable.
41	My waiting time for a response from the chatbot was short.
42	The chatbot is quick to respond.

Appendix D: BotScale (Dutch Translation)

Translation BotScale from Silderhuis and Borsci (2020):

De volgende vragen werden beantwoord op een vijf-punt Likert schaal van: sterk mee onees (1) – sterk mee eens (5)

Beantwoord de volgende stellingen op basis van je ervaring met de chatbot:

1	Het was duidelijk hoe ik een gesprek met de chatbot kon beginnen.
2	Het was gemakkelijk te begrijpen hoe ik een gesprek met de chatbot kon beginnen.
3	Ik vond het makkelijk om een gesprek met de chatbot te beginnen.
4	De chatbot was makkelijk bereikbaar.
5	De chatbot functie was makkelijk te ontdekken.
6	Het was makkelijk om de chatbot te vinden.
7	De communicatie met de chatbot was duidelijk.
8	Ik werd meteen op de hoogte gebracht van de informatie die de chatbot mij kan geven.
9	Het was voor mij al gauw duidelijk wat de chatbot kan.
10	Ik moest mijn invoer meerdere keren herformuleren voordat de chatbot me kon helpen.
11	Ik moest extra goed op mijn formulering letten tijdens het communiceren met de chatbot.
12	Het was makkelijk om de chatbot te vertellen wat ik wilde dat het deed.
13	De interactie met de chatbot voelde als een lopend gesprek.
14	De chatbot hield de context in het oog.
15	Het gesprek dat de chatbot voerde was relevant.
16	De chatbot leidde me naar de relevante service.
17	De chatbot gebruikte hyperlinks om me naar mijn doel te leiden.
18	De chatbot kon me verwijzen naar de website of een dienst wanneer nodig.
19	De interactie met de chatbot voelde veilig met betrekking tot privacy.
20	Ik denk dat de chatbot me inlicht over mogelijke privacy problemen.
21	Ik denk dat de chatbot mijn privacy waarborgt.
22	Ik had het gevoel dat mijn intenties werden begrepen door de chatbot.
23	De chatbot begeleidde mij naar mijn doel.
24	Ik denk dat de chatbot begrijpt wat ik wil en helpt me mijn doel te bereiken.
25	De chatbot gaf tijdens het gehele gesprek relevante informatie.
26	De chatbot gaf behulpzame reacties op elk moment in het proces.

27	De chatbot gaf relevante informatie wanneer ik die nodig had.
28	De hoeveelheid informatie die ik ontving was niet te veel en niet te weinig.
29	De chatbot gaf me de juiste hoeveelheid informatie.
30	De chatbot gaf me alleen de informatie die ik nodig had.
31	De chatbot kon omgaan met situaties waarin de rode draad van het gesprek niet duidelijk was.
32	De chatbot vertelde me op een vriendelijke manier wanneer het me niet kon helpen.
33	Als de chatbot op een probleem stuitte, reageerde het op gepaste wijze.
34	Ik vond de antwoorden van de chatbot duidelijk.
35	De chatbot gaf alleen begrijpelijke antwoorden.
36	De antwoorden van de chatbot waren gemakkelijk te begrijpen.
37	Ik had het gevoel dat de antwoorden van de chatbot klopten.
38	Ik denk dat de chatbot alleen betrouwbare informatie geeft.
39	De informatie die de chatbot gaf leek betrouwbaar en juist.
40	De reactietijd van de chatbot was redelijk.
41	Ik hoefde kort te wachten op een antwoord van de chatbot.
42	De chatbot reageerde snel.

Appendix E: Dutch translation SUISQ-R

1. Ik zal waarschijnlijk dit systeem weer gebruiken
2. Ik voelde mij zelfverzekerd terwijl ik dit systeem gebruikte
3. Ik kon zonder enige moeite vinden wat ik nodig had
4. Het systeem liet mij voelen alsof ik de controle had
5. Het systeem maakte gebruik van alledaagse taal
6. Het systeem leek beleefd
7. Het systeem sprak mij op een professionele manier aan
8. Het systeem leek vriendelijk
9. De stem van het systeem klonk als een gewoon persoon
10. De stem van het systeem klonk natuurlijk
11. De stem van het systeem klonk enthousiast en/of opgewekt
12. Ik had het gevoel alsof het systeem te lang sprak voordat ik kon reageren
13. In de berichten zat veel herhaling
14. Het systeem was te spraakzaam

Appendix F: Chatbots and tasks

1. https://www.ato.gov.au/

Voer de volgende taak uit met de chatbot in het Engels:

Je bent recentelijk vanuit Nederland naar Australië verhuisd. Je wilt weten wanneer de deadline is om je belastingaangifte te doen en gebruikt de ATO chatbot om meer te weten te komen.

English description:

You moved to Australia from the Netherlands recently. You want to know when the deadline is to lodge/submit your tax return using ATO's chatbot to find out.

2. https://www.amtrak.com/home

Voer de volgende taak uit met de chatbot in het Engels:

U wilt graag met de trein reizen van Boston naar Washington D.C. terwijl u in de Verenigde Staten bent. U wilt de chatbot van Amtrak gebruiken om de kortst mogelijke reis te boeken op 8 oktober. Uw vertek station is Back Bay Station.

English description:

You would like to travel from Boston to Washington D.C. while being in the USA. You want to use Amtrak's chatbot to book the shortest trip possible on the 8th October. Your departure station is Back Bay Station.

3. https://www.inbenta.com/en/

Binnenkort heeft u een sollicitatiegesprek bij Inbenta. Daarom wilt u Inbenta's chatbot gebruiken om het adres van het kantoor in Mexico te vinden.

English description:

You have an interview with Inbenta in a few days and you want to use Inbenta's chatbot to find out the address of Inbenta's Mexico office.

4. https://www.uscis.gov/

U bent een Amerikaanse staatsburger die in het buitenland woont en wilt graag stemmen in de komende verkiezingen. U wilt de chatbot gebruiken om er achter te komen hoe dit kan.

English description:

You are a US citizen living abroad and want to vote in the upcoming federal elections. You want to use the USCIS chatbot to find out how.

5. https://www.hsbc.co.uk/ (HSBC UK)

U woont in Nederland maar gaat reizen naar Turkije voor 2 weken. Tijdens uw reis wilt u graag uw HSBC credit card kunnen gebruiken bij pin- en geldautomaten. U wilt de chatbot van HSBC gebruiken om de relevante procedure hiervoor te vinden.

English description:

You live in the Netherlands but are travelling to Turkey for 2 weeks. During your travel, you would like to be able to use your HSBC credit card overseas at payment terminals and ATMS. You want to use HSBC's chatbot to find out the relevant procedure.

6. https://www.absolut.com/en/

U wilt een fles Absolut vodka drinken met uw vrienden vanavond. Één van uw vrienden mag geen gluten consumeren. Daarom wilt u de chatbot van Absolut gebruiken om er achter te komen of Absolut Lime gluten bevat.

English description:

You want to buy a bottle of Absolut vodka to share with your friends for the evening. One of your friends cannot consume gluten. You want to use Absolut's chatbot to find out if Absolut Lime contains gluten or not.

7. https://www.emiratesholidays.com/gb_en/

U werd net wakker en realiseerde dat u de verjaardag van uw partner bent vergeten. Wanhopig probeert u een verjaardagscadeau te bedenken en uw idee is om samen op vakantie te gaan naar Parijs. U gaat naar de website van "Emirates Holidays" en gebruikt de chatbots om een vakantie te boeken van 4 tot 9 september voor 2 personen. U vertrekt vanaf het vliegveld van London Heathrow (LHR). Al het overige is niet belangrijk, sinds u vandaag het cadeau nodig heeft.

English description:

You just woke up and realize that you forgot that it's your significant other's birthday. Desperately, you are thinking about a birthday present and your idea is a holiday together in Paris. You visit the Emirates Holidays page and use Emirates Holidays' chatbot to book a holiday from the 4th September until the 9th September to Paris for two persons. Your departure airport is London Heathrow (LHR). Everything else is not important, as you just need a present for today.

8. https://www.utwente.nl/en/education/master/#world-class-research-institutes

U bent een buitenlandse student en wilt graag een Master opleiding volgen aan Universiteit Twente. Uw naam is Jack/Jacky en uw email adres is abc@def.com. Je bent geïnteresseerd in de master Nanotechnologie en wilt beginnen in September 2021. U heeft uw bachelor voltooid aan Universiteit Twente. U vraagt de chatbot wat de mogelijkheden zijn voor een studiebeurs.

English description:

You are a foreign student who would like to do a Master's degree at the University of Twente. Your name is Jack/Jacky and your Email address is abc@def.com. You are interested in doing your master

in Nanotechnology in September 2021. You did your bachelor at the Utwente in the Netherlands. You ask the Utwente chatbot what options for a scholarship are available.

9. https://seattleballooning.com/

U wilt uw moeder een ballonvaart cadeau geven voor haar verjaardag. Vooraf wilt u graag weten hoe lang een ballonvaart duurt. Daarnaast wilt u ook weten wat er gebeurt als het slecht weer is op de dag van uw vlucht. U wilt daarom de chatbot gebruiken om dit te vragen.

English description:

You want to gift your mother a balloon flight for her birthday. However, you would first like to know how long a balloon flight takes. Additionally, you would like to know what happens if there is bad weather on the day of your trip. You want to use the chatbot to ask these questions.

Appendix G: R Markdown

Available on request at dr. S. Borsci.

Appendix H: 14-item BotScale

Factor	Pres
F1 Conversation start and privacy	It was clear how to start a conversation with the chatbot.
	It was easy to find the chatbot.
	I believe the chatbot informs me of any possible privacy issues.
F2 Communication quality	Communicating with the chatbot was clear.
	It was easy to tell the chatbot what I would like it to do.
	The chatbot maintained relevant conversation.
	The chatbot guided me to the relevant service.
	I felt that my intentions were understood by the chatbot.
	The chatbot gave relevant information during the whole conversation.
	The amount of information received was neither too much nor too less.
	I found the chatbot's responses clear.
	I feel like the chatbot's responses were accurate.
F3 Graceful breakdown	The chatbot explained gracefully when it could not help me.
F4 Perceived speed	My waiting time for a response from the chatbot was short.