Bachelor Thesis

Testing of a usability assessment tool for chatbots: Investigating the effect of believing that a chatbot might be a human

Steffen Neumeister (s1977180)
s.neumeister@student.utwente.nl

University of Twente
Faculty of Behavioural, Management and Social Sciences
Department of Cognitive Psychology and Ergonomics

Examination Committee:
Dr. Simone Borsci
Prof. Dr. Frank van der Velde

June 2020

UNIVERSITY OF TWENTE.

Abstract

Information-retrieval chatbots have gained more importance for customer service in recent years, but a reliable and valid usability measure tailored towards chatbots is still missing. The Usability Satisfaction Questionnaire (USQ) might solve this issue. This study tested the reliability and validity of the USQ by comparing it to the UMUX-Lite, which has already shown high reliability and validity. Furthermore, a PCA was carried out to find the most reliable distribution of components of the USQ, and the results have been compared to a previous study. Thirty-nine participants interacted with five chatbots each and completed two tasks per chatbot. They filled out the USQ and the UMUX-LITE after each interaction. Additionally, the belief that a human controls a chatbot, as well as trust in relation to the USQ has been investigated. For this, 15 participants were deceived into believing that some chatbots are controlled by humans. The PCA suggested a six-component structure with 32 items and a positive correlation has been found between the USQ and the UMUX-LITE. Moreover, a positive correlation has been found between the USQ and trust as well as between the USQ and the belief that a human controlled the chatbot. The results suggest that the USQ is a reliable and valid measure of user satisfaction, and the extracted components overlapped with those of previous studies. However, all outcomes should be seen with caution, as Covid-19 influenced the execution of the study and the belief that a human controlled the chabot had mixed results after further analysis.

Keywords: chatbot, usability, user satisfaction, conversational agent, trust

Table of contents

Introduction

Conversational agents are programs that are trained to simulate human behavior and use natural language (Radziwill & Benton, 2017). Humans can interact with conversational agents by using speech, text, touch and more, although those two examples are the most common (McTear, 2017). The conversational agent will respond in natural language, and most of the time, the output will match the input given by a human as long as the question or query is rather simple (Gnewuch, Moran, & Maedche, 2018). One of the first conversational agents was ELIZA which was developed by Weizenbaum (Ireland, 2012). ELIZA was a program that acted like a therapist and gave according to responses (Ireland, 2012). This worked so well that Weizenbaum's secretary asked to meet ELIZA in person after the interaction (Weizenbaum, 1976).

Conversational agents were initially created in the 1960s, and their aim was quite different at that time (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019). They were used in the Turing test to test whether a participant would believe that the interaction partner was a human rather than a machine (Saygin, Cicekli, & Akman, 2000). To be more precise, the Turing test, proposed by Alan Turing, consists of a machine, a human and a judge/interrogator. The judge/interrogator should not be an expert and has the task to find out which interaction partner is the human and which is the machine (Saygin et al., 2000). In the field of conversational agents, the Loebner Prize Competition is an annual event where conversational agents compete in the Turing test (Bradeško & Mladenić, 2012). Until the year 2014, none of the conversational agents passed the Turing test and the winner of the Loebner Prize was the conversational agent that seems most human-like (Bradeško & Mladenić, 2012). However, in 2014, a machine was able to pass the Turing test at the Royal Society London (Warwick & Shah, 2015).

Nowadays, conversational agents are mostly used for customer service or other customer-oriented approaches (Jenkins, Churchill, Cox, & Smith, 2007; Araujo, 2018). Their use increased dramatically since 2016, and so did the quality of conversational agents (McTear, 2017). Speech-based conversational agents are called virtual or digital assistants (Gnewuch et al., 2018). However, the more commonly used version of conversational agents are chatbots who use text messages to communicate (Araujo, 2018). Their main function is information-retrieval, and chatbots make it less complicated, especially for people who are not that familiar with the internet (Gnewuch et al., 2018). While humans possess natural language as an innate capability, they have to learn how websites and designs on the internet work in order to interact with them. Additionally, chatbots are similar to existing messengers

like Facebook Messenger or WhatsApp on which many chatbots are already used (ChatBot, n.d.). Hence, an interaction using natural language is much easier for humans (Gnewuch et al., 2018). Another advantage of chatbots used for customer service is that they reduce the amount of customer service calls which means that companies can save money by having a good chatbot. Overall, chatbots have many advantages compared to traditional interfaces and "participants of Ciechanowski et al.'s study (2019) expected more frequent usage of conversational agents in the future" (Boecker & Borsci, 2019).

As chatbots already exist for almost 60 years and their use, as well as popularity, rose over time, research has also dealt with the technology and analyzed it. Here, user trust has been identified as an important determinant for the success of chatbots. There is often a skepticism towards new technologies, and this is also the case for chatbots (Araujo, 2018; Trivedi, 2019). Luo, Tong, Fang, and Qu (2019) showed that customer purchase rates decreased by 79.7% when the customer was informed that the interaction partner is a chatbot beforehand. One reason for that might be that users have more demanding expectations regarding efficiency and rationality than they have towards humans. To be more precise, humans expect that chatbots behave human-like, possess faster and more accurate information-processing abilities, produce high output, and use appropriate language (Kim, Park, & Kim, 2003; Jenkins et al., 2007). However, this also means that people may trust computerized systems more than other humans if the system actually has a high quality (Przegalinska et al., 2019).

There are three dimensions of trust that one can distinguish regarding chatbots, namely, ability/expertise, privacy/safety and anthropomorphization (Przegalinska, Ciechanowski, Stroz, Gloor, & Mazurek, 2019).

- Expertise and ability to describe whether the user thinks that the chatbot is able to solve the user's problem. Ability is evaluated by performance factors like conversation length or customer retention while expertise depends on credibility or the user's perception of the chatbot's expertise (Przegalinska et al., 2019).
- Privacy/safety is important, especially, when the query is about sensitive topics like health (Przegalinska et al., 2019). If the user is not convinced that his/her data is safe, the user will probably not engage with the chatbot.
- Anthropomorphization can increase trust in chatbots (Przegalinska et al., 2019). As chatbots are human-like, humans tend to build trust relationships with them. This is supported by several studies which have shown that users enjoyed interacting with a chatbot (Weizenbaum, 1976; Ciechanowski et al., 2019).

All three dimensions of trust influence the user's satisfaction of the interaction with chatbots. However, it should be noted that anthropomorphization only works when the chatbot is actually human-like because people put more trust in anthropomorphized systems compared to mindless ones (Przegalinska et al., 2019).

Another factor that influences trust in chatbots is the uncanny valley effect. This effect occurs when a technology appears more human-like and results in discomfort and uneasiness (Mori, 1970). Likeability decreases the more human-like a technology appears, but it increases again when technology reaches the state of being perfectly human-like (Mori, 1970). Furthermore, the uncanny valley is mostly related to artificial human faces, so it does not apply to text-based chatbots without an avatar face. However, it may affect the likeability of chatbots with a human face negatively (Ciechanowski et al., 2019).

Apart from investigating trust, past research focused on designing and developing chatbots that imitate humans as closely as possible. Even though a large emphasis was put on design and creating human-like chatbots, previous studies pointed out that there are still design problems and chatbots are often lacking a convincing and engaging way of interacting with users (Jenkins et al., 2007; Mimoun, Poncin, & Garnier, 2012; Gnewuch et al., 2018). The problem is that creating chatbots is done without taking the end-user into account (Shackel, 2009). In the end, one might have a chatbot that can communicate quite well, has a large range of functions, and deceives the user into believing that it is a human. However, all of that is useless if the chatbot is not able to meet the user's needs in a satisfying way. Due to the fact that many chatbots fail to satisfy the user, more research needs to be done in order to assess the usability of chatbots.

Here, several general measures of perceived usability already exist. For example, SUS (Brooke, 1996), UMUX (Bosley, 2013), UMUX-LITE (Lewis, Utesch & Maher, 2013) and CSUQ (Lewis, 2002) are all standardized measures which have shown high validity and reliability in different contexts (Lewis, 2018; Borsci, Federici, Bacci, Gnaldi & Bartolucci, 2015). They are called general measures of perceived usability because they can be applied in different contexts with different technologies, and they are all questionnaires that ask about the subjectively perceived satisfaction of the user.

According to Balaji and Borsci (2019), there are several reasons why these general measures are not as effective as a more specific evaluation tool that is tailored to chatbots. First of all, a more specific measure enables chatbot developers to improve concrete parts of their chatbot rather than puzzling why the overall usability is low. This was also confirmed by Tariverdiyeva and Borsci (2019) who used the UMUX-Lite to measure perceived usability.

Secondly, there are major differences between traditional interfaces and chatbot interfaces. In most cases, the chatbot is waiting for the input given by the user while giving no or just a few hints of what it expects. A traditional interface differs in terms of transparency, as the features are presented in a specific way that enables the user to get to know its functions and limits.

Overall, there is a gap in research concerning the usability assessment of chatbots and their application in customer service. Boecker and Borsci (2019) and Balaji and Borsci (2019) tried to reduce this gap by developing a questionnaire based on a study from Tariverdiyeva and Borsci (2019). The Usability Satisfaction Questionnaire (USQ) consists of 42 items measuring user satisfaction after interacting with a chatbot. In their studies, participants interacted with five chatbots by doing an information-retrieval task and filled out the USQ after each interaction. Boecker and Borsci (2019) identified several important factors to be the most determining such as “general usability”, “ease of getting started”, “perceived privacy and security”, “response time” and “articulateness”. However, they also indicated that the USQ and the results of their studies need further validation on a larger scale. Moreover, Balaji and Borsci (2019) stated that evaluating a chatbot based on one task, which sometimes only takes a few seconds, may not be enough.

As chatbots are increasingly used in customer service and generally, for information-retrieval, there is also an increasing need for a reliable and valid usability assessment tool. Therefore, we will replicate and extend the previous work of Boecker and Borsci (2019) and Balaji and Borsci (2019).

Aims of this study

The main goal of this study is to investigate the most reliable distribution of the factors of the USQ. We expect that the results will mostly mirror the results of these two previous studies. However, slight changes will also be expected due to the fact that there are partly different chatbots used, and there are also differences in the study design. Next, the UMUX-LITE is a general measure for perceived usability, so it will be tested whether the results of the USQ and the UMUX-LITE correlate. It is expected that they do correlate because this was also the case in the studies of Balaji and Borsci (2019) and Boecker and Borsci (2019). *Research question 1. Combined, these two analyses will be used to answer the research question of how reliably and validly the USQ measures user satisfaction concerning chatbots.*

Additionally, we will investigate whether thinking that the chatbot might be a human has an effect on the user’s evaluation. *Research question 2. Hence, we will try to answer to what extent the belief that a human controls the chatbot influences the USQ scores.* For this, an additional item will be used that asks to what extent the interaction partner was perceived

to be a human. If consumer's purchase rates after interacting with a chatbot decrease by 79.7% when the chatbot is identified as such beforehand, (Luo et al., 2019) it is expected that the USQ score increases when a user thinks the interaction partner is a human. Furthermore, this is expected due to less demanding expectations and the effect of anthropomorphization.

Lastly, the effect of trust on the USQ score will be analyzed. Humans are often skeptical about new technologies like chatbots and have higher expectations in comparison to human operators. This suggests that trust can have a significant influence on the evaluation of a chatbot regardless of its quality. The first two dimensions of trust are already covered to a certain degree by some items of the USQ, namely, trust in relation to expertise or ability and in relation to privacy/safety (Przegalinska et al., 2019). *Research question 3. Consequently, we will explore to what extent trust and the USQ are associated.* Here, we will compare the USQ with an item asking about the trustworthiness of a chatbot before and after the interaction. As it has been identified as one of the most important determinants for the success of chatbots, we assume that trust and the USQ scores correlate strongly.

Methods

Participants

The BMS Test Subject Pool system SONA and convenience sampling were used to recruit 15 participants (M_{age} = 25.87 years, SD_{age} = 13.11 years) consisting of eight females and seven males. However, the data of participants from a partner study that used the USQ as well (Dehmel & Borsci, 2020) were included to get a total of 39 participants (M_{age} = 25.77 years, SD_{age} = 10.87 years) who evaluated five chatbots each. Nobody participated in both studies, so all 39 participants were unique. Combined, there have been 19 males and 20 females with nationalities of German (N = 30), Dutch (N = 6), German-Dutch (N = 1), English (N = 1) and French (N = 1). The designs of these studies were slightly different, and the complete data was neither used to investigate the effect of believing that a chatbot is a human nor the importance of trust. The eligibility for both studies was restricted to participants above the age of 18 years with a sufficient understanding of the English language. Students who registered via SONA received a reward of two credits in the BMS Test Subject Pool system for their participation.

Procedure and materials

In total, eleven chatbots were used consisting of six chatbots that were already assessed by Boecker and Borsci (2019) and Balaji and Borsci (2019) while we added five new chatbots without any prior user satisfaction evaluation. Each participant interacted with five

chatbots, and the allocation of chatbots per participant was randomized. There were two tasks prepared for each chatbot which the participant completed by interacting with the chatbot.

While the study started in a face-to-face context, we had to switch to an online design using Skype as a matter of communicating with the participants. This was due to the fact that the Covid-19 pandemic limited personal contact. Furthermore, the Booking.com chatbot was sorted out due to its limited functions during the Corona crisis and replaced by the ManyChat chatbot. The study roughly took between one hour and 90 minutes per participant. First of all, a short oral introduction to the study was given. Here, a script was used to make sure that the procedure for each participant was similar (Appendix A). Afterwards, there was time to carefully read the informed consent (Appendix B) and indicate whether the conditions of the study were acceptable (Appendix E). It should be noted that the informed consent switched from a paper to an online version due to the Corona crisis (Appendix C). The participant was told that in some cases, a human controls the chatbot if it is not well enough developed. The participants should have therefore indicated to what extent they had the impression they were interacting with a human rather than a bot. We called this the “type of interaction partner” item (TIP). This seems similar to the Turing test, but there are major differences. None of the chatbots was controlled by humans, and the task was to figure out whether a machine is actually controlled by a human rather than the other way around.

After informed consent, the participant’s demographic information was collected. Then, a hyperlink to the first chatbot was given. The participant was able to look at the chatbot for a quick moment and answered the pre-interaction trust item based on this first impression. Next, the tasks were done by interacting with the chatbot. An item asking about the task difficulty was given after each task, (Sauro & Dumas 2009) and the USQ had to be filled out after completing both tasks. The USQ consists of 42 items which the participants answered after each interaction with a chatbot. Then, the UMUX-LITE with two items (Lewis et al., 2013) had to be answered. Lastly, the post-interaction trust item and the TIP item were presented. This whole procedure was repeated for the remaining four chatbots. Furthermore, all participants have been biased after interacting with the first chatbot where no bias occurred yet. For the second and third chatbot, it was told that they are rather good chatbots meaning that there is a lower probability that a human will step in. On the other hand, it was told that the fourth and fifth chatbot are rather bad chatbots meaning that there is a much higher probability that a human will step in. As the choice and order of the chatbots presented were random, this information was not true and used to deceive the participant. Afterwards, we revealed that all interaction partners were chatbots without human interference in addition to

the fact that the quality of the chatbot indicated was deception as well. Here, the participant had the opportunity to state that the data should not be used due to this deception. This was done using a debriefing sheet which was changed from a paper version to an online version like the informed consent (Appendix D). The SONA points would have been given regardless of the participants' choice. Finally, the participant was thanked for the participation, and we made sure that all necessary information in case of further questions or remarks about the research were provided.

The software Qualtrics was run to administer the USQ consisting of the 42 items, the UMUX-LITE (Lewis et al., 2013), the task difficulty item (Sauro & Dumas, 2009), the TIP item and a pre- and post-trust item. Additionally, informed consent forms or debriefing sheets were used both before and after the study which was changed from paper to online.

Data Analysis

First, a check for outliers using graphs was used. Then, normality of the data was tested using the Shapiro-Wilk test. After that, descriptive statistics were calculated for each scale. The UMUX-LITE (Lewis et al., 2013) has two items with an overall score ranging from 0 to 100. USQ consists of 42 items with a five-point Likert scale resulting in a minimum score of 42 and a maximum score of 210. However, items 10 and 11 were negatively recoded, as they were also negatively formulated. The TIP item has a five-point Likert scale as well while both the pre- and post-trust item have a scale from 0 to 100 each. Lastly, the task difficulty item has a raw score from 1 to 10 where 1 means that the task is considered to be very difficult and 10 means that it is considered to be very easy (Sauro & Dumas, 2009). For further analysis, the variables were rescaled to intervals ranging from 0 to 1 to harmonize the scales. Furthermore, descriptive statistics for the USQ scores of several smaller study populations within the whole study population were calculated to compare them. To be more precise, the study design (remotely or in person) and the study pools (this study or Dehmel and Borsci (2020)) have been compared. The participants of Dehmel and Borsci (2020) were referred to as "Dataset 1" while the participants from this study were referred to as "Dataset 2".

Next, a principal component analysis was carried out and the model assumptions were checked. Here, the Kaiser-Meyer-Olkin Criterion (KMO) and Bartlett's test of sphericity were computed. Then, Kaiser's criterion, which demands an eigenvalue greater than one, and the scree plot were used to determine the number of extracted components. Furthermore, an oblique rotation was used. The inclusion criteria for a component was that there are at least three items with a factor loading of .5 or higher, although exceptions were possible if it

seemed reasonable. Further analyses regarding the reliability of the scale were performed by computing Cronbach's alpha for each component.

Besides, the correlation between the total scores of the USQ and the UMUX-LITE scores were computed. For the correlation, 97.5 % confidence intervals were calculated using bootstrapping with 9999 replicates of the correlation estimate.

The same was done for the total score of the USQ and the TIP item. Again, bootstrapping with 9999 replicates was used to compute 97.5 % confidence intervals.

Next, the correlation between trust and the USQ scores was computed for both the pre- and post-trust item. The same bootstrapping conditions were used to do this.

Then, a One-Way ANOVA has been performed for the different bias conditions to check their influence on the USQ scores.

Lastly, the correlation between the USQ scores and the mean of the task difficulty for both tasks was computed using bootstrapping with 9999 replicates and 97.5 % confidence intervals. This was done to check the dependency of task difficulty for the USQ.

Results

Outliers and descriptive statistics

First of all, the data for the Booking.com chatbot of one participant has been deleted, because the chatbot stopped working due to the Corona crisis. Hence, 194 responses have been taken into account for the analysis. No outliers have been detected for both the USQ scores and the UMUX-LITE. However, the Shapiro-Wilk test revealed that the data for both of them is not normally distributed. Next, the rescaled scores of the USQ ranged from .32 to .92 (M = .67, SD = .15) whereas the rescaled scores of the UMUX-LITE ranged from .13 to 1.00 (M = .72, SD = .25). Lastly, the rescaled scores of the TIP had a range from .00 to 1.00 (M = .09, SD = .19). Furthermore, the descriptive statistics for the different study designs (remotely or in person) and study pools (Dataset 1 or Dataset 2) can be seen in table 1.

Table 1

Descriptive statistics for different study pools/designs

Type	N	Minimum	Maximum	Mean	Std. Deviation
USQDataset1	120	.32	.90	.68	.15
USQDataset2	74	.36	.92	.67	.14
USQ-test in Person	55	.32	.90	.64	.14

USQ-test remotely	139	.32	.92	.69	.15

Principal component analysis

A principal component analysis was conducted for the 42 items of the USQ using oblique rotation. The sampling adequacy was verified by the Kaiser-Meyer-Olkin Criterion, KMO = .88. Furthermore, Bartlett's test of sphericity was statistically significant with x^2 (861) = 5517.23, $p < .01$. Next, Kaiser's criterion suggested a ten-component solution based on eigenvalues that are greater than one. This solution explained 72.08 % of the variance, and it was also confirmed by the scree plot (figure 1). Hence, it was decided to extract ten components.

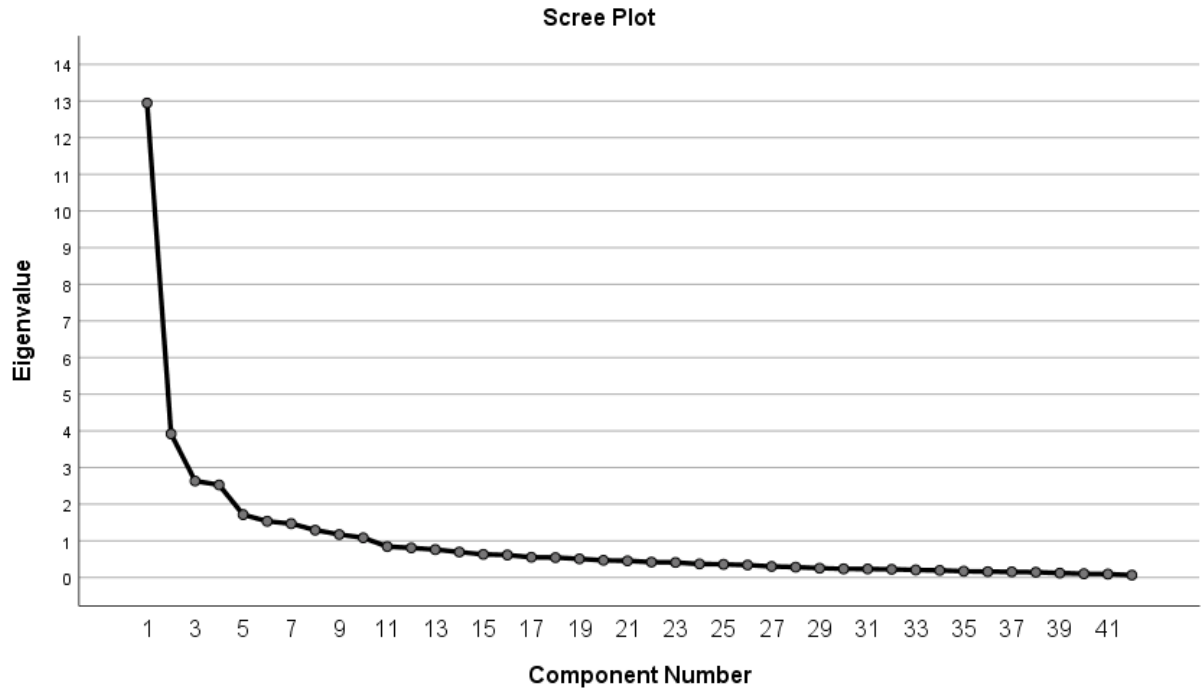

Figure 1. Scree plot of the PCA for 42 items

However, the pattern matrix showed that several components had only two (or one) items with factor loadings above .5, namely, components five, six, eight, nine and ten. Those were removed from the analysis except one component who had two items with a factor loading above .5 and one item with a factor loading of .49. As a clear-cut exclusion point does not make sense, it was decided to keep this component resulting in a six-component solution (Appendix H). This solution explained 60.34 % of the variance.

The items with a factor loading below .5 were removed with some exceptions. To be more precise, items 7, 8, 9, 12, 15, 17, 18, 22, 23 and 24 were taken out while items 10 and 11 were kept even though they formed a component consisting of only two items. This was due to the fact that they are about the topic of rephrasing which is quite important for chatbots according to the researcher's experience after performing the study. For example, the chatbot of the University of Twente asks about the nationality of the user at some point. If the answer was "China", the chatbot understood this. However, if the answer was "Chinese", the chatbot was not able to understand this. Similar examples can be found for almost all of the chatbots that were used in this study. As item 10 and 11 are the only ones asking about this, it was decided to keep them. Going on with reliability analyses, all components had a Cronbach's alpha of $\alpha > .7$ except component five with $\alpha = .68$, which is still close to the satisfactory score of $\alpha = .7$. Therefore, no further actions have been taken regarding reliability. Consequently, the result is a six-component structure with 32 items explaining 65.29 % of the variance (Table 2).

Table 2

Oblique rotated factor loadings for six components and 32 items[a]

Item	F1	F2	F3	F4	F5	F6
USQ38	**.81**					
USQ28	**.79**					
USQ29	**.79**					
USQ39	**.75**					
USQ25	**.71**					.,33
USQ35	**.70**					
USQ30	**.69**					
USQ34	**.69**					
USQ36	**.65**					-.31
USQ37	**.64**					
USQ26	**.57**					
USQ27	**.57**					.33
USQ16	**.51**					
USQ1		**.85**				
USQ5		**.83**				

USQ4		**.82**				
USQ6		**.81**				
USQ2		**.79**				
USQ3		**.71**				
USQ41			**.97**			
USQ42			**.95**			
USQ40			**.91**			
USQ19				**.87**		
USQ21				**.85**		
USQ20				**.78**		
USQ32					**.72**	
USQ33					**.68**	
USQ13					**.64**	
USQ14					**.58**	
USQ31					**.52**	
USQ10						**.74**
USQ11						**.73**
Eigenvalues	**9.30**	**3.70**	**2.51**	**2.34**	**1.61**	**1.44**
% of Variance	**29.06**	**11.56**	**7.84**	**7.32**	**5.01**	**4.50**

[a] factor loadings > .3 supressed

The first component consisting of items 16, 25, 26, 27, 28, 29, 30, 34, 35, 36, 37, 38 and 39 was labelled *general satisfaction* (Table 3). Components two to four were quite similar compared to the structure extracted by Boecker and Borsci (2019) and therefore, received the same names. Here, component two, consisting of items 1 to 6 was called *ease of getting started*. Moreover, component three with items 40 to 42 was labelled *response time*. Next, component four consisting of items 19 to 21 received the label *perceived privacy and security*. Component five with items 13, 14, 31, 32 and 33 was named *keeping track of context*. Lastly, component six with items 10 and 11 received the label *flexibility of linguistic input*.

Table 3

Labels of components[a]

Component	Item	Feature
1: General satisfaction	16 The chatbot guided me to the relevant service.	Reference to service
	25 The chatbot gave relevant information during the whole conversation	Maxim of relation
	26 The chatbot is good at providing me with a helpful response at any point of the process.	Maxim of relation
	27 The chatbot provided relevant information as and when I needed it.	Maxim of relation
	28 The amount of received information was neither too much nor too less	Maxim of quantity
	29 The chatbot gives me the appropriate amount of information	Maxim of quantity
	30 The chatbot only gives me the information I need	Maxim of quantity
	34 I found the chatbot's responses clear.	Understandability
	35 The chatbot only states understandable answers.	Understandability
	36 The chatbot's responses were easy to understand.	Understandability
	37 I feel like the chatbot's responses were accurate.	Perceived credibility

	38 I believe that the chatbot only states reliable information.	Perceived credibility
	39 It appeared that the chatbot provided accurate and reliable information.	Perceived credibility
2: Ease of getting started	1 It was clear how to start a conversation with the chatbot.	Ease of starting a conversation
	2 It was easy for me to understand how to start the interaction with the chatbot.	Ease of starting a conversation
	3 I find it easy to start a conversation with the chatbot.	Ease of starting a conversation
	4 The chatbot was easy to access.	Visibility
	5 The chatbot function was easily detectable.	Visibility
	6 It was easy to find the chatbot.	Visibility
3: Response time	40 The time of the response was reasonable.	Response time
	41 My waiting time for a response from the chatbot was short.	Response time
	42 The chatbot is quick to respond.	Response time
4: Perceived	19 The interaction with the chatbot felt secure in terms of privacy.	Perceived privacy and security

privacy and security		
	20 I believe the chatbot informs me of any possible privacy issues.	Perceived privacy and security
	21 I believe that this chatbot maintains my privacy.	Perceived privacy and security
5: Keeping track of context	13 The interaction with the chatbot felt like an ongoing conversation.	Ability to maintain themed discussion
	14 The chatbot was able to keep track of context.	Ability to maintain themed discussion
	31 The chatbot could handle situations in which the line of conversation was not clear	Graceful responses in unexpected situations
	32 The chatbot explained gracefully when it could not help me	Graceful responses in unexpected situations
	33 When the chatbot encountered a problem, it responded appropriately	Graceful responses in unexpected situations
6: Flexibility of linguistic input	10 I had to rephrase my input multiple times for the chatbot to be able to help me.	Flexibility of linguistic input
	11 I had to pay special attention regarding my phrasing when communicating with the chatbot.	Flexibility of linguistic input

[a] labels mostly taken from Boecker and Borsci (2019)

Relationship between USQ and UMUX-LITE

The data were not normally distributed, therefore, we used Kendall's Tau to calculate the correlations. Here, the USQ and the UMUX-LITE correlated with $r = .71$, $p < .01$. Furthermore, bootstrapping with 9999 replicates and 97.5% confidence resulted in [.65, .76]. A visualization of this relationship can be seen in figure 2.

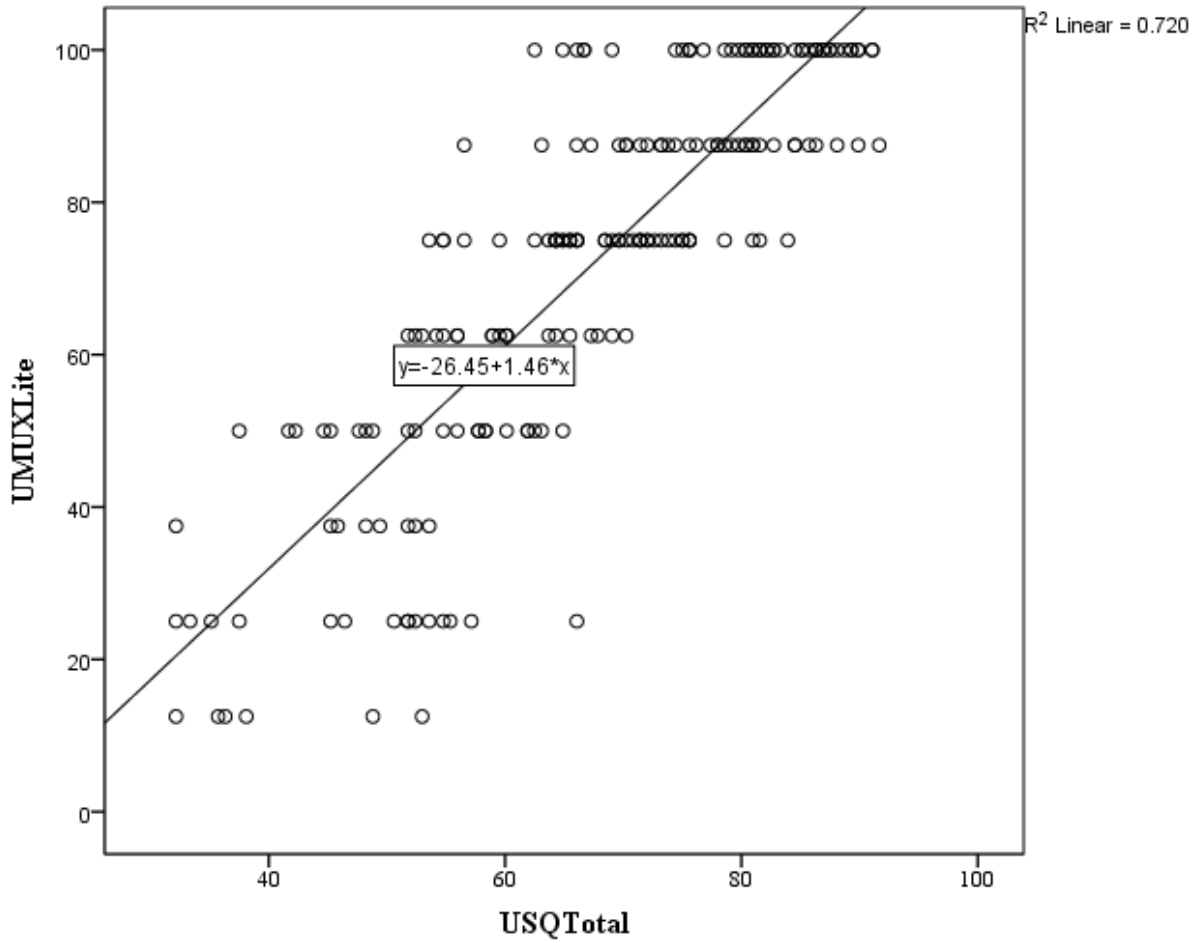

Figure 2. The scatterplot in percentages with UMUX-LITE as the dependent and USQ as the independent variable

Relationship between USQ and TIP

For this comparison, only the 74 responses from "Dataset 2" were taken into account, as this is one of the items that was not included in the study of Dehmel and Borsci (2020). Again, Kendall's Tau was used, and the USQ correlated with the type of interaction partner item with $r = .24$, $p < .05$. The bootstrapping with 9999 replicates and 97.5% confidence resulted in [.02, .43]. Figure 3 shows that there have been mainly values of 0% for the TIP which means that the participant never had the feeling that the chatbot was controlled by a human and only five which were above 25%. Furthermore, doing the same analysis with the

UMUX-LITE which has already been proven to be reliable, shows no significant effects between the TIP and the UMUX-Lite.

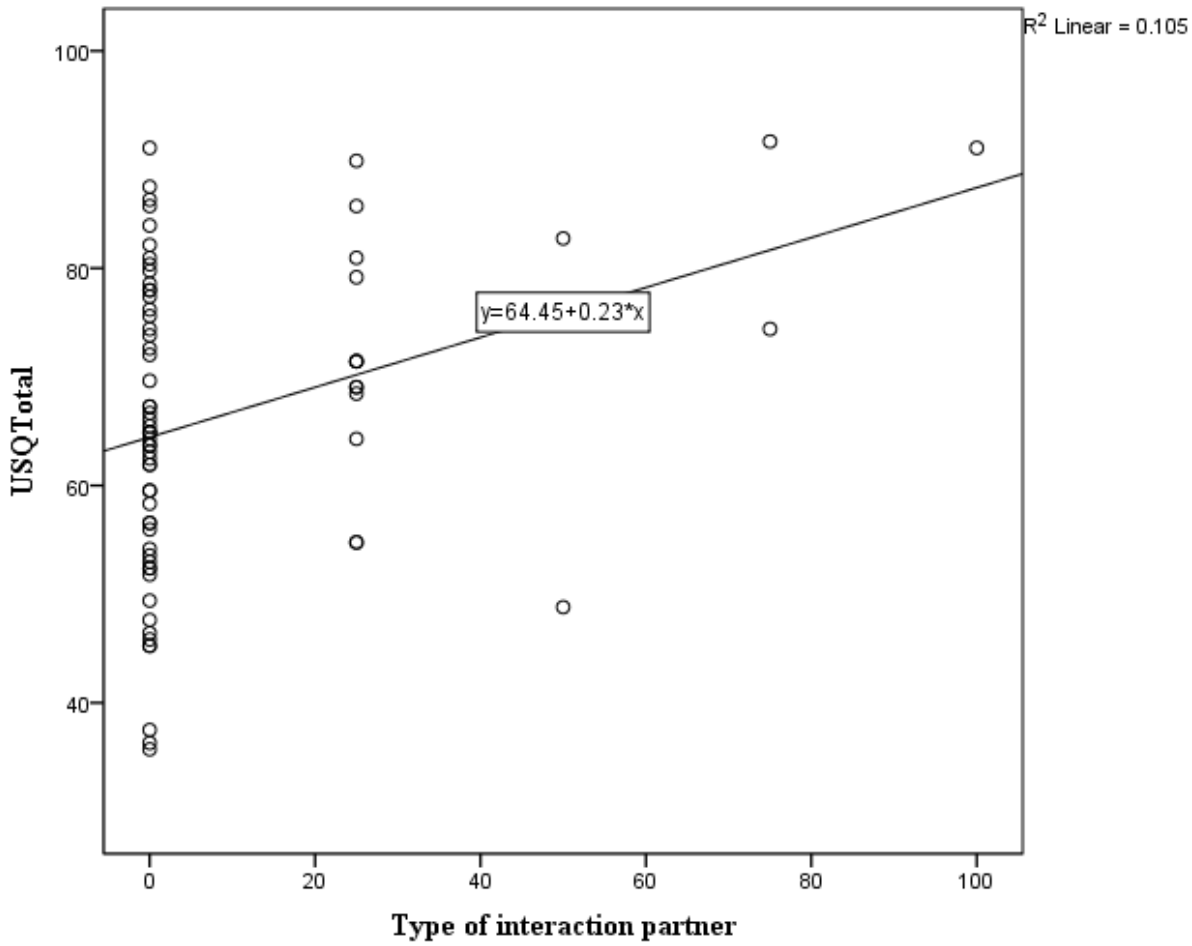

Figure 3. The scatterplot in percentages with USQ as the dependent and TIP as the independent variable

Relationship between USQ and trust

Kendall's Tau was used for 74 responses to compute this relationship with $r = .37$, $p < .01$ for the pre-trust item and $r = .22$, $p < .01$ for the post-trust item. Using bootstrapping with 9999 replicates, the 97.5% confidence intervals have a range of [.19, .53] for the pre-trust item and [.04, .39] for the post-trust item. Figure 4 shows the scatterplot of these relationships.

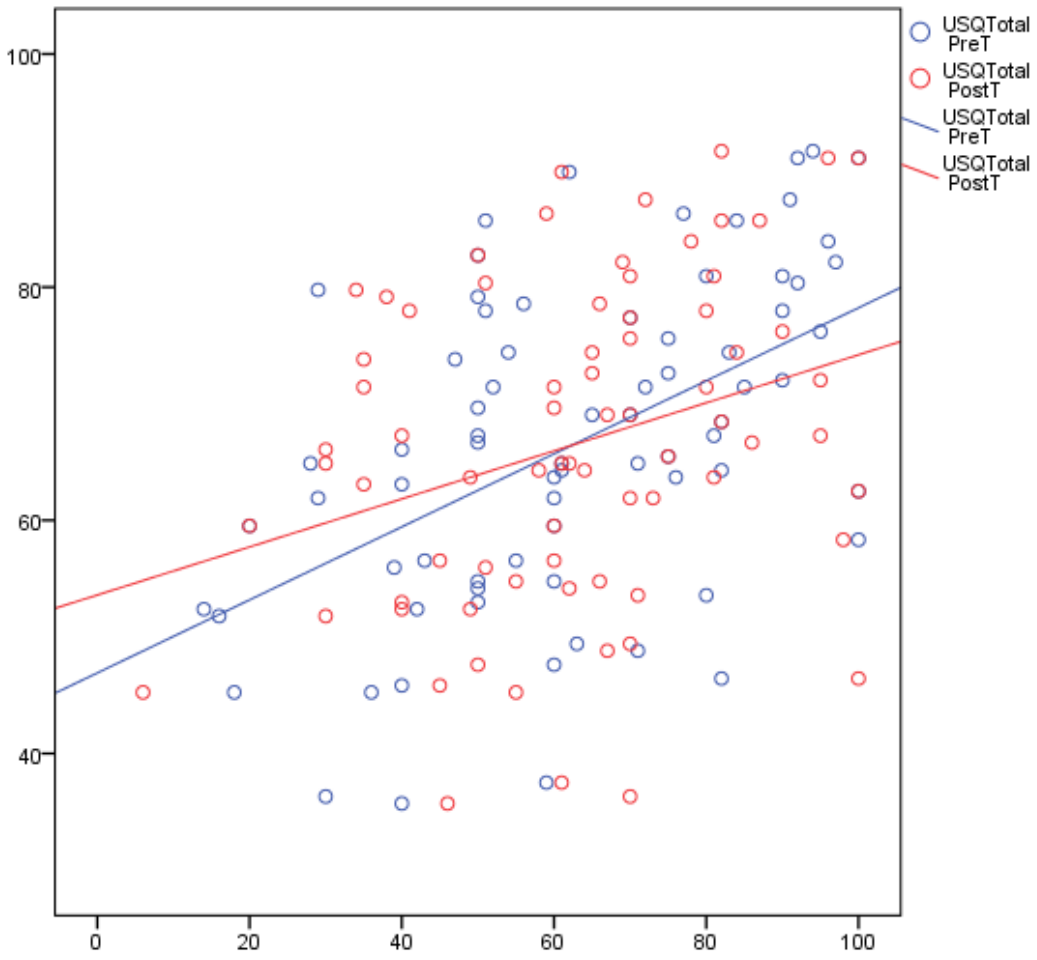

Figure 4. The scatterplot in percentages with USQ as the dependent and both pre- and post-trust as independent variables

Comparison of bias conditions

The three conditions “no bias”, “bias that the chatbot is good” and “bias that the chatbot is bad” have been compared to each other using a One-Way ANOVA with 74 responses. A visualization of the USQ scores in each bias condition can be found in figure 5. There were no statistically significant differences between group means as determined by the One-Way ANOVA (F (2,71) = 2.81, p = .07). As the analysis could not find meaningful differences in the scores and the boxplots show much overlap, further post-hoc tests were not used.

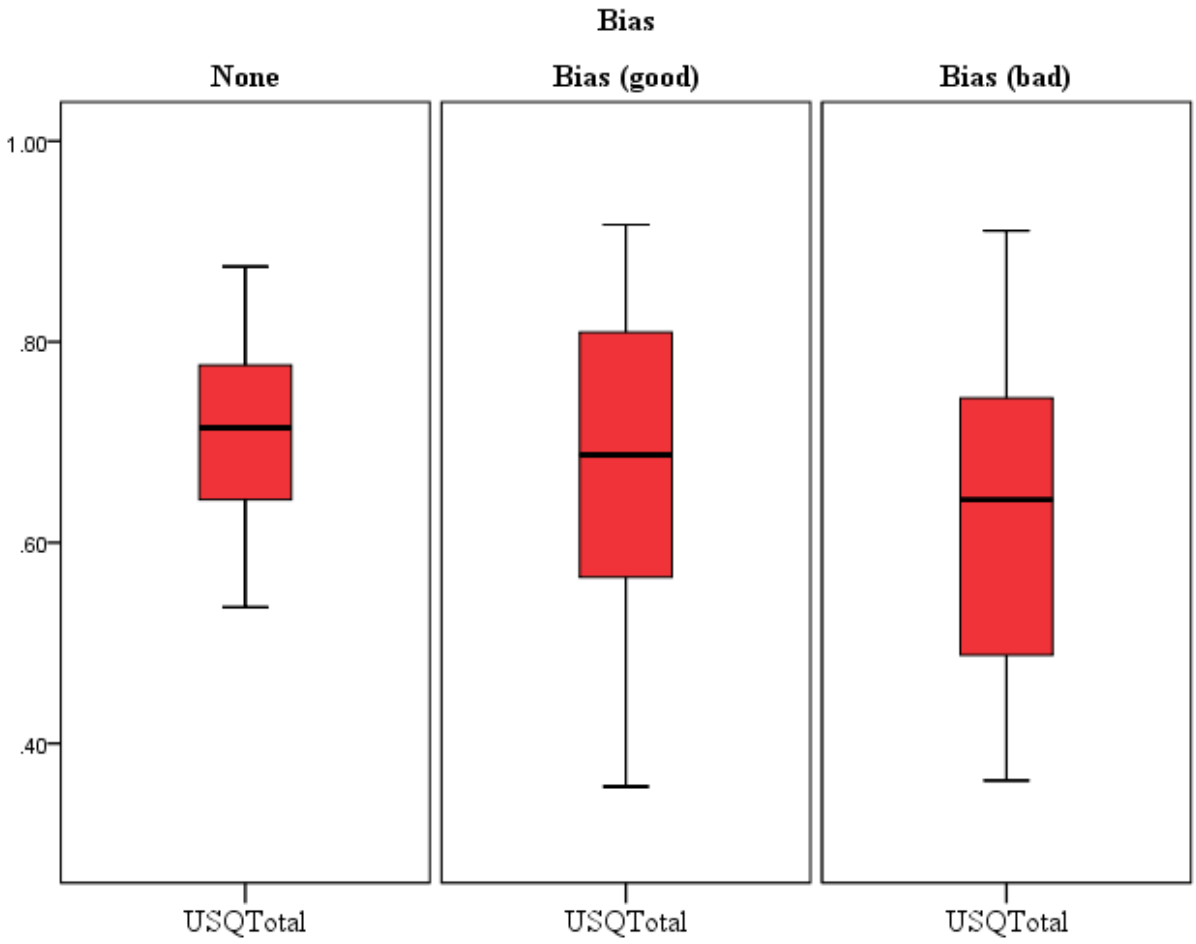

Figure 5. Boxplots for the USQ scores in each bias condition

Relationship between USQ and task difficulty

For this analysis, all 194 responses have been taken into account and Kendall's Tau was used again. The scores correlated with $r = .43$, $p < .01$. Next, the bootstrapping with 9999 replicates and 97.5% confidence resulted in [.33, .52]. The visualization of this can be seen in figure 6. Here, it should be noted that task difficulty scale was reversed meaning that a higher value indicates that the task was considered to be easier and vice versa.

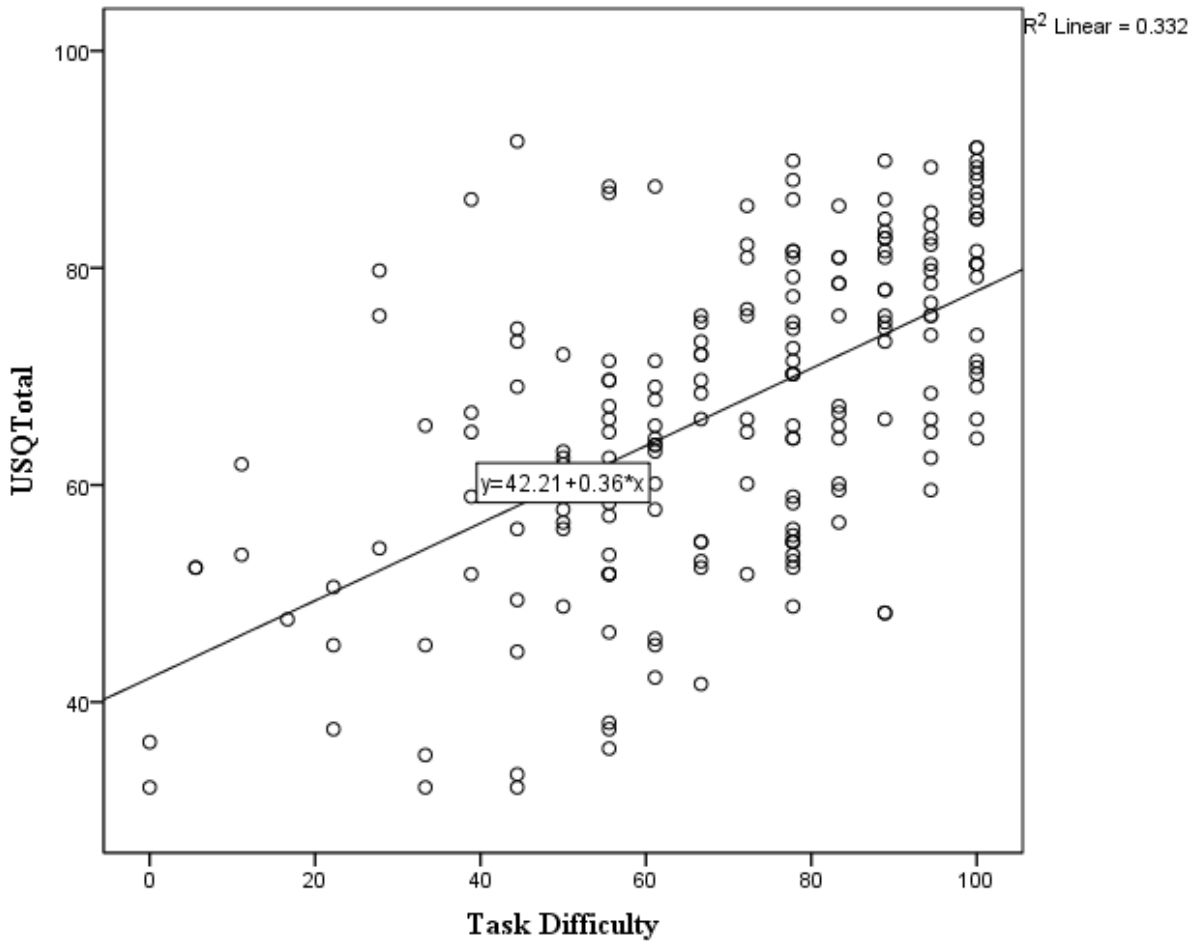

Figure 6. The scatterplot in percentages with USQ as the dependent and task difficulty as the independent variable

Discussion

This study tried to replicate and extend the studies of Boecker and Borsci (2019) and Balaji and Borsci (2019) using the USQ to assess the user satisfaction of chatbots. However, the Coronavirus Covid-19 interfered and forced us to change part of the study design as well as the scope of this study. Furthermore, a similar study (Dehmel & Borsci, 2020) was merged with this study to investigate the reliability and validity of the USQ. Additionally, the effect that a user thinks that a chatbot is a human was analyzed using a deception.

The first research question asks how reliably and valid the USQ measures user satisfaction concerning chatbots. To answer this, a principal component analysis has been carried out in addition to computing the correlation between the USQ and the UMUX-LITE. Although there are differences in the results of the principal component analysis between this study and the previous study of Boecker and Borsci (2019), several extracted components are quite similar. One should also note that this study coded items 10 and 11 of the USQ

negatively resulting in a logical difference between both studies and the creation of the component *flexibility of linguistic input*. Moreover, different exclusion criteria were used to arrive at the final component structure, which meant that fewer items were excluded in this study. Nonetheless, this study mirrored most of the results from Boecker and Borsci (2019) with slight changes as expected.

These combined results suggest that *ease of getting started, perceived privacy and security* as well as *response time* are key features that the USQ measures in relation to chatbots. Furthermore, both studies showed several items that are not related to a single specific feature, but rather indicate *general satisfaction*. However, only four out of 13 from these items overlap between both studies which could be seen as a problem and a sign that the reliability is questionable even though Cronbach's alpha suggests the contrary in almost all cases. Lastly, the feature *keeping track of context* differs from the feature articulateness that was extracted in Boecker's and Borsci's (2019) study. Here, items 32 and 33 were included in both studies and item 31 which was only included in our study is related to the same topic than these two items. The major differences are item 13 and 14, which go more into the direction of context rather than understandability. As the reliability for this component was slightly below the desired value, it is still questionable whether it fits. Overall, the rather similar component structures indicate that this is a quite reliable distribution of components, especially, for the ones that are exactly the same.

Next, the relationship between the USQ and the already reliable and valid UMUX-LITE has been affirmed which was also the result of Boecker and Borsci (2019) and Balaji and Borsci (2019). This suggests that USQ is a reliable and valid measure of user satisfaction with chatbots.

The second research question asked to what extent the belief that a human controls the chatbot influences the USQ scores. Here, the belief that a chatbot is a human had a rather small, but significant effect on the USQ scores. Interestingly, this effect was not confirmed by the UMUX-LITE, which correlates with the USQ. Hence, the results should be seen with caution, especially because the correlation between this belief and the USQ score is quite low. Moreover, there have been many answers of "never" (did I believe that the chatbot was a human) meaning that there is only little variance. When the participant believed that it was a human to some extent, the evaluation was usually higher, but this is only based on a handful of answers. On the other hand, there is much variation when the participant thought that it was never a human. This stands in conflict with the results of Luo et al. (2019) who showed that pre-identification of a chatbot as such reduces the likelihood that the user purchases a product

suggesting that the user is also less satisfied. However, one has to take into account that the study designs are quite different, and our study did not try to replicate the previously used study design. Overall, the results regarding this topic are quite mixed and need further investigation.

The third research question asked to what extent trust and the USQ are associated. This study affirmed the assumption that trust is a major indicator of user satisfaction. It also seems like trust before interacting with the chatbot correlates more with user satisfaction than the trust estimation after the interaction. Looking back at figure 4, it is visible that there is much variation within the results which might contradict with the assumptions to a certain extent. There might be several explanations that could account for these results. One could be the uncanny valley effect which might have influenced the likeability for some chatbots with a human face (Ciechanowski et al., 2019). Another reason might be that trust is also dependent on the perceived expertise of the chatbot which could not be assessed beforehand (Przegalinska et al., 2019). However, this would mean that the correlation of the post-trust item should be higher than the one of the pre-trust item. Nevertheless, this study showed the importance of trust concerning chatbots with a certain ambiguity regarding the interpretation of the results.

Moving on, we could not find meaningful differences between the three bias conditions which can be considered as something positive, because it means that the USQ scores were not majorly influenced by the bias given. An explanation for this could be that it is questionable to what extent one or two sentences used for the bias affect the participant if the interaction with the chatbot that took much longer felt entirely different.

Lastly, the evaluation of the chatbots and hence, also the USQ is highly task-dependent. The difficulty as well as the phrasing and other factors related to the task can have a major influence on the evaluation of the chatbot. This was also confirmed by the correlational analysis that showed that the USQ score was usually higher when the task was perceived to be easier. However, one should also take into account that completing a task might be more difficult when the quality of the chatbot is worse. Hence, the results make sense and increase the reliability of the USQ even more. Nevertheless, one or two tasks can almost never cover the whole scope of a chatbot which means that the creation of suitable tasks is quite important as well. Without good tasks, the USQ may have limited usefulness for the evaluation of user satisfaction with chatbots.

Limitations

There are five major topics regarding weaknesses of this study, but also of the USQ. First of all, the design had to be changed to Skype calls which might have had an effect on the evaluation of the chatbots. While this was not extensively investigated, at least the mean scores between both study designs differed. Related to that, the researcher was not always able to see the screen of the participant, which also led to some difficulties and misunderstandings.

Next, the results of two studies have been merged together to create the component structure, compare USQ with UMUX-LITE and USQ with task difficulty. While the study designs were similar, different sub-aspects were analyzed and the procedure was not exactly the same. This could have resulted in slight differences in the evaluation even though the mean scores of both studies are quite similar.

Going on, this study tried to simulate real-world conditions as much as possible. Consequently, almost no hints were given on what to do to accomplish the task. While task clarification was always provided if requested or needed, the participant was not told how to solve a task or when it was completed. This was based on the participants' intuition to mirror a real-life situation as closely as possible. The only exceptions to this rule were tasks where the chatbot asked, for example, about personal details, so it was told whether the task was completed or not. However, even without these exceptions, there have been multiple instances where a participant thought that a task was completed, although it was not and vice versa. A real user would know what he/she is looking for and when an answer of the chatbot is satisfactory while our participants had to imagine they were looking for that information. Hence, another variable that might have influenced the evaluation is the imagination of the participant and the extent to which an answer is seen as satisfactory.

Next, the belief that a chatbot is a human was only measured using one single item. This could have been done in a more extensive way. Related to that, all chatbots were actual chatbots requiring a deception of the participant. Here, one could use an actual human to act like a chatbot and create a similar condition like the Turing test. Additionally, while anthropomorphization was given as a phenomenon that may influence an evaluation, it is questionable to what extent it was applicable when considering the study design. This is due to the fact that the participant was told that a human might step in and take control over the chatbot. However, anthropomorphization refers more to the idea of humanizing a machine rather than thinking that a machine is controlled by a human. Furthermore, it only works when the chatbot actually behaves human-like, which requires a high quality (Przegalinska et al., 2019).

Lastly, the order in which the bias was given was always the same. This could have affected the evaluation as well because the participant was already familiar with some chatbots at the end.

Future directions

One weakness that could be fixed by a future study is that the last component of the extracted six-component structure has only two items. These two items are negatively formulated, so a future study could add one or two positively formulated items. This might solve the two-item problem, and one could compare the items in order to check whether the participant was actually reading the questions rather than blindly answering all of them.

If a future study tries to replicate what we have done, the researchers should consider providing a clean environment where the chatbot is presented apart from the website. The advantage of this mostly blank page is that the participant will not be distracted by other features of a website.

Also, the bias given should rather be reversed if used again in a future study. It was told that a chatbot is bad meaning that there is a higher probability that a human will step in. This was done because it made sense along with the explanation used to deceive the participant. However, the results suggest that this is the other way around, so another explanation should be used to say that a better chatbot is probably just a human. Additionally, it could be considered whether already tested chatbots should be classified into a bias based on their quality rather than just randomly picking chatbots without considering their quality. One advantage of this is that a possibly used deception would be more believable.

While this study focused on quantitative aspects regarding the USQ, a follow-up study could also analyze the interaction with a chatbot in terms of qualitative aspects. It was already mentioned that participants did not always complete the task even though they occasionally thought they did. A future study could analyze the video recordings of the interaction and for instance, classify it in terms of task completion. To be more precise, whether the task was completed and whether the participant thought it was completed. Of course, there are even more aspects that a future study could analyze regarding quality. The reason why this might be useful is to test to what extent these studies represent a real-life situation where a user knows whether an answer is satisfactory.

Lastly, the effect of trust could be further analyzed, as the results of this study were somewhat ambiguous. Possible fields of interest are the relation to privacy and safety or the importance of a first impression, as trust before the interaction correlated more with the USQ in this study.

Conclusion

Overall, this study affirmed the reliability and validity of the USQ for measuring user satisfaction concerning chatbots. Furthermore, the study showed that trust and task difficulty are also related to the evaluation of a chatbot. However, the results regarding the influence of believing that a chatbot might be a human are mixed and need further clarification.

The most important limitation of this study is the Coronavirus Covid-19 which reduced the number of participants and forced a transition to Skype calls. Therefore, it is questionable whether the outcomes are replicable under normal conditions.

References

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:10.1016/j.chb.2018.03.051

Balaji, D., & Borsci, S. (2019). *Assessing User Satisfaction with Information Chatbots: A Preliminary Investigation* (Master's thesis). University of Twente, Netherlands.

Boecker, N., & Borsci, S. (2019). *Usability of information-retrieval chatbots and the effects of avatars on trust* (Bachelor's thesis). University of Twente, Netherlands.

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.

Bosley, J. J. (2013). Creating a short usability metric for user experience (UMUX) scale. *Interacting with Computers, 25*(4), 317-319.

Bradeško, L., & Mladenić, D. (2012). A survey of chatbot systems through a Loebner prize competition. In *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies* (pp. 34-37).

Brooke, J. (1996). SUS - A quick and dirty usability scale. In *Usability evaluation in industry* (pp. 189-194). London: Taylor and Francis.

ChatBot. (n.d.). *Facebook Messenger chatbot integration*. Retrieved June 11, 2020, from https://www.chatbot.com/integrations/facebook-messenger/

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems, 92*, 539-548. doi:10.1016/j.future.2018.01.055

Dehmel, A., & Borsci, S. (2020). *On the usefulness of the preliminary usability satisfaction questionnaire (USQ), its dimensionality, and the impact of user characteristics* (Bachelor's thesis). University of Twente, Netherlands.

Gnewuch, U., Morana, S., & Maedche, A. (2018). Towards designing cooperative and social conversational agents for customer service. In *ICIS 2017: Transforming Society with Digital Innovation*. Retrieved from https://www.researchgate.net/publication/320015931_Towards_Designing_Cooperative_and_Social_Conversational_Agents_for_Customer_Service

Ireland, C. (2012). *Alan Turing at 100*. Retrieved from

https://news.harvard.edu/gazette/story/2012/09/alan-turing-at-100/

Jenkins, M., Churchill, R., Cox, S., & Smith, D. (2007). Analysis of user interaction with service oriented chatbot systems. *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments Lecture Notes in Computer Science,* 76-83. doi:10.1007/978-3-540-73110-8_9

Kim, B., Park, K., & Kim, J. (2003). Satisfying different customer groups for IS outsourcing: A korean IS company's experience. *Asia Pacific Journal of Marketing and Logistics,15*(3), 48–69. doi:10.1108/13555850310765006

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction, 14*(3-4), 463-488.

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction, 34*(12), 1148-1156.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of CHI 2013* (pp. 2099–2102). Paris, France: ACM. doi:10.1145/2470654.2481287

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*. doi:10.1287/mksc.2019.1192

McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *Lecture Notes in Computer Science Future and Emerging Trends in Language Technology. Machine Learning and Big Data,*38-49. doi:10.1007/978-3-319-69365-1_3

Mimoun, M. S., Poncin, I., & Garnier, M. (2012). Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services, 19*(6), 605-612. doi:10.1016/j.jretconser.2012.07.006

Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33-35.

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:10.1016/j.bushor.2019.08.005

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability

questionnaires. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*. doi:10.1145/1518701.1518946

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and machines*, *10*(4), 463-518. doi:10.1023/A:1011288000451

Shackel, B. (2009). Usability–context, framework, definition, design and evaluation. *Interacting with computers*, *21*(5-6), 339-346. doi:10.1016/j.intcom.2009.04.007

Tariverdiyeva, G., & Borsci, S. (2019). *Chatbot's perceived usability in information retrieval tasks: an exploratory analysis* (Master's thesis). University of Twente, Netherlands.

Trivedi, J. (2019). Examining the Customer Experience of Using Banking Chatbots and Its Impact on Brand Love: The Moderating Role of Perceived Risk. *Journal of Internet Commerce,*1-21. doi:10.1080/15332861.2019.1567188

Warwick, K., & Shah, H. (2015). Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(6), 989-1007. doi:10.1080/0952813x.2015.1055826

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman & Co.

Appendix A

Script

<<For researcher only: enter participant code>>

Welcome to our study. We appreciate you helping us out today! We are in the process of testing a measure to assess user satisfaction with information-retrieval chatbots. Today, you will be testing some chatbots and providing us with your feedback by responding to questionnaires. You will be presented with five chatbots, each with two associated tasks to do. After using each chatbot, you will have a few questionnaires to respond to through an online survey software.

In some cases, companies develop a chatbot and realize that it only has limited capacities. Hence, they employ a human to act like a chatbot until the actual chatbot is good enough and replaces the human's answers. We know when a human is used to step in, and we have designed our tasks so that a person will step in when it is needed. This does not necessarily mean that as soon as the human steps in, you will only interact with the person going onwards.

Please focus on achieving the tasks. At the end of each interaction, we will ask you to what extent you had the feeling that you chatted with a person. The session is expected to last for no more than one and a half hours. We would like to record your voice and the screen for data analysis purposes. If you are not okay with this, please let us know. There are more details in the informed consent which you must read and sign before proceeding further.

<<Give participant informed consent form>>

First, please fill in the demographic questionnaire.

You will now begin testing chatbots. Each provided task is a short realistic scenario – you, as the participant, should try your best to imagine yourself in those situations i.e. imagine that you're looking for that information for the first time. If you do not understand the situation or task, let me know. Once you feel like you have achieved the task, or if you feel that the task is not achievable, please let me know.

Remember that we aim to assess the quality of the chatbot not you, if you cannot do something it's not your fault, but there is a problem with the tool. Also remember that there is no wrong or right answer in this experiment, we are interested in what you think about the chatbot.

Your behaviour and responses will help us understand how users will use these chatbots.

Do you have any questions? Are you ready to start? If so, you may begin with the first chatbot. Follow the instructions on the screen and if you

have questions, you may ask me.

<<Start recording the screen>>

<<First chatbot: no new information>>

Now, the next two chatbots are considered good chatbots meaning that there is a much lower probability that a human will step in.

<<Chatbot 2 and 3 are tested>>

On the other hand, the last two chatbots are considered as bad chatbots meaning that there is a much higher probability that a human will step in.

<<Chatbot 4 and 5 are tested>>

Appendix B

Participant Information Sheet

Title: Validating a measure of user satisfaction for evaluating interactions with chatbots
Principal investigator: Steffen Neumeister
Co-investigator/ Supervisor: Dr Simone Borsci

Before you decide to take part in this study, it is important for us that you understand why the research is being done and what it will involve. Please take the time to read the following information carefully and then decide whether or not you would like to take part. The researchers can be contacted if there is anything you wish to clarify.

Purpose of the study

This study aims to validate a new measure for evaluating user satisfaction with chatbot interaction. It means that we are trying to test whether the questionnaire you will fill out reliably measures user satisfaction after interacting with a chatbot. Furthermore, we would like to know if participants are able to detect if the chatbot is actually a human. In some cases, companies let humans take control over the chatbot if the chatbot is not well developed enough or if the chatbot has problems dealing with the customer.

Your role as participant

Note that your participation is entirely voluntary. Refusal or withdrawal will involve no penalty, now or in the future. If you wish to withdraw yourself from the study at any point of the session, please simply inform the responsible researcher. Involvement in this study is not related to any risks of physical or mental kind for you as the participant.

You are asked to perform a usability test on several chatbots using the developed measurement tool. The experiment is including you to perform certain tasks in a chatbot when asked. Afterwards, you will have to fill in the questionnaire developed for usability testing of information-retrieval chatbots. You should also indicate how much you had the impression that the interaction partner was a human rather than a chatbot.

Personal data

Personal information, namely age, gender, nationality and educational/professional background will be collected for demographic purposes.

Videotaping and Questionnaire

When performing the usability testing, each participant’s questionnaire data will be anonymized and securely stored for our research team to analyse. Additionally, each screen will be videotaped while performing usability testing with each chatbot and will capture the participant’s actions as they perform the tasks. These video recordings will enable the research team to retrieve valuable information about how users perceive and interact with chatbots. Additionally, the participant will be audio recorded. All data will be made

anonymous before stored and secured on a separate hard drive to which the research team and supervisor will have access during the research period while writing the bachelor thesis. When data evaluation is finished, the access will belong solely to the supervisor. The research has the potential to be published and therefore, the data will have a retention period of approximately 12 months, when it is expected to be published. During the retention period, only the supervisor will have access to it.

Ethical review of the study

The project has been reviewed and approved by the International Review Board.

Contact details

Principal Researcher
Steffen Neumeister
s.neumeister@student.utwente.nl

Co-Investigator/ Supervisor
Dr. Simone Borsci
s.borsci@utwente.nl

Contact Information for Questions about Your Rights as a Research Participant

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-bms@utwente.nl

Appendix C

Consent form (only for Skype)

	Yes	No
I have read and understood the study information, 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 questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves a video- and audio-recorded usability session. I am aware that my face, voice and if possible the screen will be recorded, and that this data will be treated with discretion until destroyed.	○	○
I understand that information I provide will be used for data analysis while writing a bachelor thesis and for potential 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.	○	○

Consent for recording

	Audio	Face	Screen (if possible)
I agree to be recorded	☐	☐	☐

Appendix D

Participant Debriefing Sheet (only before Covid-19)

Title: Validating a measure of user satisfaction for evaluating interactions with chatbots

Principal investigator: Steffen Neumeister

Co-investigator/ Supervisor: Dr Simone Borsci

You just completed the study. However, not everything that I have told you is correct. You were asked to indicate if the interaction partner was a chatbot or a human. The truth is that all of them were chatbots and none of them was a human. We actually cannot be fully sure if some of the chatbots were controlled by humans, because only the companies monitoring the chatbots could know this. Additionally, the information which chatbots were good and which ones were bad was completely invented. I sincerely apologize for this deception.

Purpose of the deception

It was used to make you believe that you might interact with a human. We would like to investigate whether such a bias would influence your evaluation of the chatbots. This might help us in improving the questionnaire to evaluate chatbots.

Your rights

You can withdraw your initial consent without any consequences or justification. Your data will be deleted immediately. You can also decide that you only want the video/audio recording to be deleted. In that case, your answers to the questionnaire will still be used. If you want to confirm your initial consent and let me use all the data you agreed on, you just have to sign this form without making any changes to the consent form.

Signatures

_____ _____ _____

Name of participant Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____ _____

Researcher name Signature Date

Appendix E

Consent form (only before Covid-19)

Consent Form for Validating a measure of user satisfaction for evaluating interactions with chatbots

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information dated [DD/MM/YYYY], 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 questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves a video- and audio-recorded usability session. I am aware that my face and voice will be recorded, and that this data will be treated with discretion until destroyed.	○	○
Use of the information in the study		
I understand that information I provide will be used for data analysis while writing a bachelor thesis and for potential 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.	○	○
Consent to be Audio/video Recorded		
I agree to be audio/video (only screen) recorded.	○	○

Signatures

_____ _____ _____

Name of participant Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____ _____

Researcher name Signature Date

Appendix F

Qualtrics survey flow

Only for the first chatbot the whole flow of the questionnaire is presented here, for every following chatbot the same questions were shown in the same order. Each participant was confronted with five out of the ten chatbots (11 in total, but only 10 at the same time), determined by randomization. The Booking.com chatbot was later replaced by the ManyChat chatbot. The informed consent was excluded to avoid repetition, but it is marked where it appeared. Furthermore, the informed consent, debriefing and E-Mail were added for the Skype call participants.

Chatbots_UT - 2020BSc

Start of Block: ID

Participant ID

_____

End of Block: ID

Start of Block: Informed Consent

As shown in Appendices B and C

End of Block: Informed Consent

Start of Block: Demographics

Gender

▼ Male (1) ... Prefer not to say (3)

Age

_____

Nationality

- ○ Dutch (4)
- ○ German (5)
- ○ If other, please specify: (6) _____

Field of study

- ○ Psychology (4)
- ○ Communication science (5)
- ○ If other, please specify: (6) _____

	Extremely familiar (1)	Very familiar (2)	Moderately familiar (3)	Slightly familiar (4)	Not familiar at all (5)
How familiar are you with chatbots and/or other conversational interfaces? (1)	○	○	○	○	○

	Definitely yes (1)	Probably (2)	Unsure (3)	Probably not (4)	Definitely not (5)
Have you used a chatbot or a conversational interface before? (1)	○	○	○	○	○

Display This Question:

If = Definitely yes

Or = Probably

Or = Unsure

	Daily (1)	4 - 6 times a week (2)	2 - 3 times a week (3)	Once a week (4)	Rarely (5)	Never (6)
How often do you use it? (1)	○	○	○	○	○	○

End of Block: Demographics

Start of Block: Emirates Holidays

Chatbot: Emirates Holidays

The chatbot can be found at: https://www.emiratesholidays.com/gb_en/

Please access the chatbot now.

Page Break

	0 10 20 30 40 50 60 70 80 90 100
On a scale from 1 to 100, how trustworthy does this chatbot appear to you?	

Page Break

Please do the following task with this chatbot.

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.

Page Break

On a scale of 1 (very difficult) to 10 (very easy), how easy did you find this task?

	1	2	3	4	5	6	7	8	9	10	
	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)	
Very difficult	○	○	○	○	○	○	○	○	○	○	Very easy

Page Break

Please do the following task with this chatbot.

You arrived in Paris and there seems to be a problem with your hotel reservation. You try to call someone at Emirates Holiday, but it's 11pm on Friday, so you cannot reach anyone. Hence, you ask Emirates Holidays' chatbot when the customer service opens on Saturday.

Page Break

On a scale of 1 (very difficult) to 10 (very easy), how easy did you find this task?

	1	2	3	4	5	6	7	8	9	10	
	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	9 (9)	10 (10)	
Very difficult	○	○	○	○	○	○	○	○	○	○	Very easy

Page Break

Based on the chatbot you just interacted with, respond to the following statements.

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
It was clear how to start a conversation with the chatbot. (1)	○	○	○	○	○
It was easy for me to understand how to start the interaction with the chatbot. (2)	○	○	○	○	○
I find it easy to start a conversation with the chatbot. (3)	○	○	○	○	○
The chatbot was easy to access. (4)	○	○	○	○	○
The chatbot function was easily detectable. (5)	○	○	○	○	○
It was easy to find the chatbot. (6)	○	○	○	○	○
Communicating with the chatbot was clear. (7)	○	○	○	○	○
I was immediately made aware of what information the chatbot can give me. (8)	○	○	○	○	○
It is clear to me early on about what the chatbot can do. (9)	○	○	○	○	○

I had to rephrase my input multiple times for the chatbot to be able to help me. (10)	○	○	○	○	○
I had to pay special attention regarding my phrasing when communicating with the chatbot. (11)	○	○	○	○	○
It was easy to tell the chatbot what I would like it to do. (12)	○	○	○	○	○
The interaction with the chatbot felt like an ongoing conversation. (13)	○	○	○	○	○
The chatbot was able to keep track of context. (14)	○	○	○	○	○
The chatbot maintained relevant conversation. (15)	○	○	○	○	○
The chatbot guided me to the relevant service. (16)	○	○	○	○	○
The chatbot is using hyperlinks to guide me to my goal. (17)	○	○	○	○	○
The chatbot was able to make references to the website or service when appropriate. (18)	○	○	○	○	○

The interaction with the chatbot felt secure in terms of privacy. (19)	○	○	○	○	○
I believe the chatbot informs me of any possible privacy issues. (20)	○	○	○	○	○
I believe that this chatbot maintains my privacy. (21)	○	○	○	○	○
I felt that my intentions were understood by the chatbot. (22)	○	○	○	○	○
The chatbot was able to guide me to my goal. (23)	○	○	○	○	○
I find that the chatbot understands what I want and helps me achieve my goal. (24)	○	○	○	○	○
The chatbot gave relevant information during the whole conversation (25)	○	○	○	○	○
The chatbot is good at providing me with a helpful response at any point of the process. (26)	○	○	○	○	○
The chatbot provided relevant information as and when I needed it. (27)	○	○	○	○	○

The amount of received information was neither too much nor too less (28)	○	○	○	○	○
The chatbot gives me the appropriate amount of information (29)	○	○	○	○	○
The chatbot only gives me the information I need (30)	○	○	○	○	○
The chatbot could handle situations in which the line of conversation was not clear (31)	○	○	○	○	○
The chatbot explained gracefully when it could not help me (32)	○	○	○	○	○
When the chatbot encountered a problem, it responded appropriately (33)	○	○	○	○	○
I found the chatbot's responses clear. (34)	○	○	○	○	○
The chatbot only states understandable answers. (35)	○	○	○	○	○
The chatbot's responses were easy to understand. (36)	○	○	○	○	○

I feel like the chatbot's responses were accurate. (37)	○	○	○	○	○
I believe that the chatbot only states reliable information. (38)	○	○	○	○	○
It appeared that the chatbot provided accurate and reliable information. (39)	○	○	○	○	○
The time of the response was reasonable. (40)	○	○	○	○	○
My waiting time for a response from the chatbot was short. (41)	○	○	○	○	○
The chatbot is quick to respond. (42)	○	○	○	○	○

Based on the chatbot you just interacted with, respond to the following statements.

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
This system's capabilities meet my requirements. (1)	○	○	○	○	○
This system is easy to use. (2)	○	○	○	○	○

Page Break

	0 10 20 30 40 50 60 70 80 90 100
On a scale from 1 to 100, how trustworthy did this chatbot appear to you? ()	

How often did you have the impression that you were interacting with a human rather than a chatbot?

	Never (26)	Sometimes (27)	About half the time (28)	Most of the time (29)	Always (30)
Answer (1)	○	○	○	○	○

End of Block: Emirates Holidays

Start of Block: NBC News

NBC
Chatbot: NBC News
The chatbot can be found at: https://www.messenger.com/t/NBCNews

Please access the chatbot now.

NBC_Task1
Please do the following task on this chatbot.

You want to use the chatbot of NBC News to find out the most recent news regarding the environment.

NBC_Task2
Please do the following task on this chatbot.

Just out of curiosity, you are also interested in the most recent special coverage, using the chatbot of NBC News.

End of Block: NBC News

Start of Block: Amtrak

Amtrak
Chatbot: Amtrak

The chatbot can be found at: https://www.amtrak.com/home

Please access the chatbot now.

Amtrak_Task1
Please do the following task on this chatbot.

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.

Amtrak_Task2
Please do the following task on this chatbot.

You have planned a trip to the USA. You are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet an old friend for a few hours and see the city. You want to use Amtrak's chatbot to find out how much it will cost to temporarily store your luggage at the station.

End of Block: Amtrak

Start of Block: Utwente

Utwente
Chatbot: Utwente
The chatbot can be found at: https://www.utwente.nl/en/education/master/chat/

Please access the chatbot now.

Utwente_Task1
Please do the following task on this chatbot.

You are a chinese student who would like to do a Master's degree at the University of Twente. Your name is Jackie/Lin 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.

Utwente_Task2
Please do the following task on this chatbot.

You are a german student who would like to do a Master's degree at the University of Twente. Your name is Alan/Sabine and your Email address is abc@def.com. You are interested in doing your master in computer science in February 2022. You did your bachelor's at the Jacobs University in Bremen. You ask the Utwente chatbot about deadlines and the admission process.

End of Block: Utwente

Start of Block: HSBC

HSBC
Chatbot: HSBC UK
The chatbot can be found at: https://www.hsbc.co.uk/

Please access the chatbot now.

HSBC_Task1
Please do the following task on this chatbot.

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.

HSBC_Task2
Please do the following task on this chatbot

You have recently moved from Amsterdam to London and would like to know how you can change your address for your HSBC card, using the chatbot of HSBC UK.

End of Block: HSBC

Start of Block: USCIS

USCIS
Chatbot: USCIS
The chatbot can be found at: http://www.uscis.gov/emma

Please access the chatbot now.

USCIS_Task1
Please do the following task on this chatbot.

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.

USCIS_Task2
Please do the following task on this chatbot.

You are planning to take a job in the USA. Since you are not a US citizen, you want to find out more about eligibility for a US- Green Card with the help of the USCIS chatbot.

End of Block: USCIS

Start of Block: ManyChat

ManyChat
Chatbot: ManyChat
The chatbot can be found at: https://www.messenger.com/t/ManyChat

Please access the chatbot now.

ManyChat_Task1
Please do the following task on this chatbot.

You want to integrate a chatbot on your companies' website. Therefore, you want to use the ManyChat's chatbot to find video tutorials to learn the basics of ManyChat.

ManyChat_Task2
Please do the following task on this chatbot.

After using the Chatbot for a while, you are getting a little bored and want to have some fun. Let the ManyChat's chatbot tell a joke to you.

End of Block: ManyChat

Start of Block: HubSpot

HubSpot
Chatbot: HubSpot
The chatbot can be found at: https://www.hubspot.com/?survey=123

Please access the chatbot now.

HubSpot_Task1
Please do the following task on this chatbot.

You have your own company and would like to grow your business even more. A former colleague recommends you Hubspot. However, you don't want to sign up for anything (even if it's free). You use Hubspot's chatbot to purely get information and get educated without using any tools. A collection of news/articles/tips would be great.

HubSpot_Task2
Please do the following task on this chatbot.

Now, you are convinced that Hubspot can help your own business. Your focus is on improving your own customer service. Before you sign up for something, you would like to know how Hubspot can improve your customer service. You use Hubspot's chatbot to get more information about this

End of Block: HubSpot

Start of Block: ATO

ATO
Chatbot: ATO
The chatbot can be found at: http://www.ato.gov.au/

Please access the chatbot now.

ATO_Task1
Please do the following task with this chatbot.

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.

ATO_Task2
Please do the following task with this chatbot.

You are a student and are wondering whether you have to lodge a tax return using the ATO's chatbot.

End of Block: ATO

Start of Block: Booking.com

Booking
Chatbot: Booking.com
The chatbot can be found at: https://www.facebook.com/messages/t/131840030178250

Please access the chatbot now.

Booking_Task1
Please do the following task on this chatbot.

You are travelling to London from 5th July to 9th July with your family. You want to use booking.com's chatbot to find a hotel room for you, your significant other and your child in Central London that does not cost more than 500€ in total

---

Booking_Task2
Please do the following task on this chatbot.

You have to attend an important business meeting from 18th to 19th of March in Amsterdam. You therefore are looking for a place to stay in the city center of Amsterdam for not more than £200 using booking.com's chatbot.

End of Block: Booking.com

Start of Block: Absolut

Absolut
Chatbot: Absolut
The chatbot can be found at: https://www.absolut.com/en/

Please access the chatbot now.

Absolut_Task1
Please do the following task on this chatbot.

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.

Absolut_Task2
Please do the following task on this chatbot.

You want to buy a bottle of Absolut vodka for a good friend. But this friend is right now on a diet

and tries to avoid sugar. You therefore want to find information about the amount of sugar in the products of Absolut using Absolut's chatbot.

End of Block: Absolut

Start of Block: Debriefing

Participant Debriefing Sheet
Title: Validating a measure of user satisfaction for evaluating interactions with chatbots
Principal investigator: Steffen Neumeister
Co-investigator/ Supervisor: Dr Simone Borsci

You just completed the study. However, not everything that I have told you is correct. You were asked to indicate if the interaction partner was a chatbot or a human. The truth is that all of them were chatbots and none of them was a human. We actually cannot be fully sure if some of the chatbots were controlled by humans, because only the companies monitoring the chatbots could know this. Additionally, the information which chatbots were good and which ones were bad was completely invented. I sincerely apologize for this deception.

Purpose of the deception
It was used to make you believe that you might interact with a human. We would like to investigate whether such a bias would influence your evaluation of the chatbots. This might help us in improving the questionnaire to evaluate chatbots.

Your rights
You can withdraw your initial consent without any consequences or justification. Your data will be deleted immediately. You can also decide that you only want the video/audio recording to be deleted. In that case, your answers to the questionnaire will still be used. If you want to confirm your initial consent and let me use all the data you agreed on, you just have to sign this form without making any changes to the consent form.

Q143 Do you want to delete any of the recordings?

	None (1)	Face (2)	Audio (3)	Screen (if applicable) (5)
Do you want to delete any of the recordings? (1)	☐	☐	☐	☐

SIG Sign here if you want to confirm your consent

End of Block: Debriefing

Start of Block: Email

EM If you want to receive information about the results of this study, you can enter your Email address in the field below.

_____

End of Block: Email

Start of Block: End

Q92
This is the end of the session.
Thank you for participating!

End of Block: End

Appendix G

SPSS code

```
*Computing rescaled total values for USQ, UMUX, IT (which is the extent to which a person believed he/she interacted with a chatbot), PT, T (which are pre- and post-trust) and task difficulty (TD)

COMPUTE USQTotal=(MEAN(USQ1,USQ2,USQ3,USQ4,USQ5,
    USQ6,USQ7,USQ8,USQ9,USQ10,USQ11,
    USQ12,USQ13,USQ14,USQ15,USQ16,USQ17,
    USQ18,USQ19,USQ20,USQ21,USQ22,USQ23,
    USQ24,USQ25,USQ26,USQ27,USQ28,USQ29,
    USQ30,USQ31,USQ32,USQ33,USQ34,USQ35,
    USQ36,USQ37,USQ38,USQ39,USQ40,USQ41,
    USQ42) - 1) / 4.
EXECUTE.

COMPUTE UMUXLite=((UMUX1 + UMUX2 - 2) / 8).
EXECUTE.



COMPUTE ITr=(IT - 1) / 4.
EXECUTE.

COMPUTE PTr=PT / 100.
EXECUTE.

COMPUTE Tr=T / 100.
EXECUTE.

COMPUTE TDr=(MEAN(TD1, TD2) - 1) / 9.
EXECUTE.

*Outlier and normality check for USQ and UMUX-LITE

EXAMINE VARIABLES=USQTotal UMUXLite
```

```
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES EXTREME
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL
  /ID=ID.

*Descriptive statistics

DESCRIPTIVES VARIABLES=USQTotal UMUXLite ITr USQAlex USQSteffen
USQSkype USQPerson
  /STATISTICS=MEAN STDDEV MIN MAX.

*initial PCA based on Kaiser's criterion

FACTOR
  /VARIABLES  USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10
USQ11 USQ12 USQ13 USQ14 USQ15
    USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
    USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
    USQ42
  /MISSING LISTWISE
  /ANALYSIS USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10 USQ11
USQ12 USQ13 USQ14 USQ15
    USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
    USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
    USQ42
  /PRINT INITIAL CORRELATION SIG KMO EXTRACTION ROTATION
  /FORMAT SORT BLANK(.30)
  /PLOT EIGEN
  /CRITERIA MINEIGEN(1) ITERATE(100)
  /EXTRACTION PC
```

```
  /CRITERIA ITERATE(100) DELTA(0)
  /ROTATION OBLIMIN
  /METHOD=CORRELATION.
```

*2nd run with ten pre-determined factors

```
FACTOR
  /VARIABLES  USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10
USQ11 USQ12 USQ13 USQ14 USQ15
    USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
    USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
    USQ42
  /MISSING LISTWISE
  /ANALYSIS USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10 USQ11
USQ12 USQ13 USQ14 USQ15
    USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
    USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
    USQ42
  /PRINT INITIAL CORRELATION SIG KMO EXTRACTION ROTATION
  /FORMAT SORT BLANK(.30)
  /PLOT EIGEN
  /CRITERIA FACTORS(10) ITERATE(100)
  /EXTRACTION PC
  /CRITERIA ITERATE(100) DELTA(0)
  /ROTATION OBLIMIN
  /METHOD=CORRELATION.
```

*3rd run with six pre-determined factors

```
FACTOR
  /VARIABLES  USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10
USQ11 USQ12 USQ13 USQ14 USQ15
    USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
```

```
  USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
  USQ42
 /MISSING LISTWISE
 /ANALYSIS USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ7 USQ8 USQ9 USQ10 USQ11
USQ12 USQ13 USQ14 USQ15
  USQ16 USQ17 USQ18 USQ19 USQ20 USQ21 USQ22 USQ23 USQ24 USQ25 USQ26
USQ27
  USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
  USQ42
 /PRINT INITIAL CORRELATION SIG KMO EXTRACTION ROTATION
 /FORMAT SORT BLANK(.30)
 /PLOT EIGEN
 /CRITERIA FACTORS(6) ITERATE(100)
 /EXTRACTION PC
 /CRITERIA ITERATE(100) DELTA(0)
 /ROTATION OBLIMIN
 /METHOD=CORRELATION.

*final PCA with six pre-determined factors and 32 items

FACTOR
 /VARIABLES  USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ10 USQ11 USQ13 USQ14
  USQ16 USQ19 USQ20 USQ21 USQ25 USQ26 USQ27
  USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
  USQ42
 /MISSING LISTWISE
 /ANALYSIS  USQ1 USQ2 USQ3 USQ4 USQ5 USQ6 USQ10 USQ11 USQ13 USQ14
  USQ16 USQ19 USQ20 USQ21 USQ25 USQ26 USQ27
  USQ28 USQ29 USQ30 USQ31 USQ32 USQ33 USQ34 USQ35 USQ36 USQ37 USQ38
USQ39 USQ40 USQ41
  USQ42
 /PRINT INITIAL CORRELATION SIG KMO EXTRACTION ROTATION
 /FORMAT SORT BLANK(.30)
 /PLOT EIGEN
```

```
  /CRITERIA FACTORS(6) ITERATE(100)
  /EXTRACTION PC
  /CRITERIA ITERATE(100) DELTA(0)
  /ROTATION OBLIMIN
  /METHOD=CORRELATION.

*Reliability check with Cronbach's alpha for the first component (similar procedure for all others)

RELIABILITY
/VARIABLES=USQ16 USQ25 USQ26 USQ27 USQ28 USQ29 USQ30 USQ35 USQ36 USQ37 USQ38 USQ39 USQ34
/SCALE('.') ALL
/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE
/SUMMARY=TOTAL.

*Correlation matrix with bootstrapping for USQTotal and UMUXLite (other variables were analysed as well and put in a similar matrix)

BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES INPUT=USQTotal UMUXLite
  /CRITERIA CILEVEL=97.5 CITYPE=PERCENTILE  NSAMPLES=9999
  /MISSING USERMISSING=EXCLUDE.
NONPAR CORR
  /VARIABLES=USQTotal UMUXLite
  /PRINT=KENDALL TWOTAIL NOSIG
  /MISSING=PAIRWISE.

*regression analysis between USQ and ITr

BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=USQTotal INPUT=  ITr
  /CRITERIA CILEVEL=97.5 CITYPE=PERCENTILE  NSAMPLES=9999
```

```
  /MISSING USERMISSING=EXCLUDE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(97.5) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT USQTotal
  /METHOD=ENTER ITr.

*One-Way ANOVA to compare bias conditions

ONEWAY USQTotal BY Bias
  /PLOT MEANS
  /MISSING ANALYSIS.
```

Appendix H

Oblique rotated factor loadings for six components[a]

Item	F1	F2	F3	F4	F5	F6
USQ28	**.83**					
USQ29	**.83**					
USQ25	**.78**					
USQ39	**.76**					
USQ30	**.76**					
USQ38	**.74**					
USQ37	**.66**					
USQ27	**.62**					
USQ35	**.59**					
USQ26	**.56**					
USQ34	**.55**					
USQ16	**.53**				.30	
USQ18	**.49**					
USQ23	**.48**				.38	
USQ36	**.47**					
USQ24	**.46**					.39
USQ22	**.46**					.38
USQ15	**.43**					
USQ17	**.42**					-.34
USQ1		**.85**				
USQ5		**.84**				
USQ4		**.82**				
USQ6		**.82**				
USQ2		**.79**				
USQ3		**.73**				
USQ41			**.99**			
USQ42			**.96**			
USQ40			**.92**			
USQ19				**.83**		
USQ21				**.82**		
USQ20				**.78**		

USQ32					**.69**	
USQ33					**.66**	
USQ13					**.62**	
USQ14					**.53**	
USQ31					**.50**	
USQ7	.30				**.39**	
USQ9					**.36**	
USQ8					**.32**	
USQ10						**.71**
USQ11						**.69**
USQ12						**.40**
Eigenvalues	**13.09**	**3.86**	**2.58**	**2.49**	**1.71**	**1.61**
% of Variance	**31.17**	**9.19**	**6.15**	**5.92**	**4.07**	**3.84**

[a] factor loadings < .3 supressed