June 21, 2021

Testing a usability scale for chatbots: The effect of familiarity on satisfaction ratings

Bachelor thesis

Niklas Pollmann
s2153386

Faculty of Behavioural, Management, and Social sciences, University of Twente

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

Abstract

Although the use of chatbots as information retrieval assistants becomes more common nowadays, there are many problems that companies encounter when implementing these tools. To find out which facets of those chatbots need improvement to better accommodate the needs of the user, usability scales for chatbots are the most effective tools available. This study aims to perform confirmatory factor analysis on the 15-item Chatbot Satisfaction Scale developed by Borsci et al. (under review). The data of 56 participants, who used an English and a German scale to rate their interaction with 10 different chatbots, were used. The confirmatory factorial analysis suggested a good model fit for the Chatbot Satisfaction Scale that could be further improved by adding a new factor and deleting two items that assessed the same concept as another item. The resulting scale composed of 13 items demonstrated good reliability ($\alpha = 0.93$) and showed a strong correlation with the UMUX-Lite. The German translation of the new scale was found to have a good reliability ($\alpha = 0.95$) that was comparable to the original English version. While a moderate to high correlation was found between the two versions, a significant difference suggested that the German version fails to measure the same concepts. Moreover, we also investigated the potential effect of familiarity on the new scale, nevertheless, results suggested no significant effects of familiarity on satisfaction after the interaction with chatbots.

Keywords: chatbot, usability, user satisfaction, Chatbot Satisfaction Scale, UMUX-LITE, familiarity

Table of contents

Introduction 3
History and development of chatbots 3
Problems in the use of chatbots 5
Characteristics of chatbots 5
Familiarity with the use of chatbots 7
Quality of use 8
The present study 9
Methods 10
Participants 10
Materials 11
Tasks 11
Procedure 12
Data Analysis 13
Results 14
Chatbot Satisfaction Scale 15
Test for normality 15
Manipulation check for supervised and unsupervised participants 16
Confirmatory factor analysis 16
Comparing the Chatbot Satisfaction Scale with 13 items and the UMUX-Lite 24
Comparing English and German version of the CSS-13 25
Effect of familiarity on the CSS-13 25
Discussion 27
Psychometric properties of the CSS-13 27
Translation of the CSS-13 and familiarity 29
Limitations of the present study and future work 30
Conclusion 32
References 33
Appendices 39
Appendix A: Five factor structure of the 15-item Chatbot Satisfaction Scale 39
Appendix B: Chatbot Satisfaction Scale (English original) 39
Appendix C: Chatbot Satisfaction Scale (German translation) 40
Appendix D: UMUX-Lite (German translation) 41
Appendix E: Chatbots and tasks 41
Appendix F: Informed consent form 44
Appendix G: Questions about familiarity 45

Introduction

If you have a smartphone or frequently use the internet, the chances are high that you have interacted with a conversational agent at some point. For example, if you recently contacted customer service it is possible that you were not conversing with another person. Instead, it may have been an artificial intelligence that is supposed to simulate human behaviour, or in other words a conversational agent (Radziwill & Benton, 2017). These virtual agents can imitate human behaviour as they are able to interactively exchange information in the form of natural language with humans (Przegalinska, Ciechanowksi, Stroz, Gloor, & Mazurek, 2019). Furthermore, conversational agents work as assistants to the user, as they engage in goal-directed behaviour by performing one or more commands after receiving the natural language input (Radziwill & Benton, 2017). A popular example of this is conversational agents such as Apple Siri, Amazon Alexa, or Google Assistant (McTear, Callejas, & Girol, 2016). This kind of conversational agents that uses speech as input are also called virtual or digital agents (Gnewuch, Morana, & Maedche, 2017). The more common use of conversational agents, however, is based a text-based input and are called chatbots (Araujo, 2018; Gnewuch, Morana, & Maedche, 2017).

History and development of chatbots

The first time conversational agents were developed was during the 1960s when they were used in the Turing test to see whether people would notice if they interacted with a computer instead of another human being (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019). The objective of these tests was to find out whether computers were able to display behaviour that is not distinguishable from the behaviour of humans. Shortly after, ELIZA was developed. This chatbot was known for using pattern matching to create the illusion of understanding towards the user (Weizenbaum, 1966). This was made possible by a script that provided the chatbot with rules on how to reply to the users' input. Advancement in the development of chatbots was made in 2014 when Microsoft published XiaoIce, an empathetic chatbot that was able to identify the emotional needs of its users (Zhou, Goa, Li, & Shum, 2020). By sending encouraging messages to the user, XiaoIce was also able to offer engaging interpersonal communication and thus satisfy the human

need for communication, affection, and social belonging (Zhou, Goa, Li, & Shum, 2020). This was enabled through a switch in the development of conversational agents from a rule-based approach to a neural-learning approach. Nowadays, most conversational agents are developed with an approach based on neural learning (Pamungkas, 2018). That is, the conversational agent is given access to datasets of recorded conversations, for example from Twitter or certain messenger services, which are then analysed by the computer to learn how to react appropriately in all different kinds of situations (Pamungkas, 2018).

Since advancements have been made in the development of artificial intelligence, chatbots have been given more attention than before (Følstad & Brandtzaeg, 2017). Another development that should be considered when looking at the rising popularity of chatbots is the increased use of instant messenger services over the last years that has helped users to become more familiar with this kind of communication (Gnewuch et al., 2017; McTear et al., 2016). As already indicated by the XiaoIce conversational agent which put emphasis on empathy, the object of the design of the chatbot became the conversation itself (Zhou et al., 2020; Følstad & Brandtzaeg, 2018). Thus, the use of chatbots will not be restricted to being a tool, but the use as a dialogue partner or assistant, as in the case of Amazon Alexa for example, will increase in the future (Huang et al., 2008). Nevertheless, currently, the most frequent use of chatbots is for Q&A and customer support (Baravesco et al., 2020). These chatbots are service-oriented and thus are designed to help customers find information on websites (Jenkins, Churchill, Cox, & Smith, 2007). Businesses benefit from chatbots in a way that they take over the work of employees working in customer service. Thereby, customers more frequently get in contact with chatbots instead of calling customer service. Businesses are thus able to significantly reduce the costs for customer service (Capgemini, 2019). Not only are chatbots cheaper than employees, but they also offer many more benefits to businesses. While employees are restricted by working hours, chatbots can operate any time of the day without the need for a break (Somasundaram, Kant, Rawat, & Maheshwari, 2019). Hence, chatbots can offer instant assistance at any time. Moreover, chatbots also have an advantage over human employees in that they can communicate with multiple customers at once, whereas a customer service employee can only interact with one person at a time. This often results in waiting times that are eliminated with the introduction of chatbots which can reply to any

number of customers instantly (Somasundaram et al., 2019). Currently, chatbots are used by only 14% of all Dutch companies (van Os, Hachmang, Akpinar, Kreuning, & Derksen, 2018). Nevertheless, another 47% of Dutch companies planned to implement chatbots over the course of the next two years (van Os et al., 2018).

Problems in the use of chatbots

Despite the advancements in the development of chatbots and their rise in popularity, many companies that use chatbots still encounter problems. Many customers still prefer to interact with humans and are still sceptical about the new technology (Araujo, 2018). Oftentimes customers perceive information as too personal to be shared with a chatbot (Zamora, 2017). As can be seen in these examples, trust plays an important role in the interaction with chatbots (Corritore, Kracher, & Wiedenbeck, 2003). The fact that purchase rates decreased when customers were informed that they were interacting with a chatbot only affirms the need to consider such aspects in the development of chatbots (Luo, Tong, Fang, & Qu, 2019). Yet, many chatbots are still designed without consideration of the needs users may have (Shackel, 2009). Another problem arises as the demands users have towards technology such as chatbots is much higher than for other human beings (Seeger, Pfeiffer, & Heinzl, 2017). Commonly, there is the expectation that chatbots can process information much faster and in a more accurate manner compared to humans. Accordingly, many users expect chatbots to save them time, however, that is often not the case (Zamora, 2017). Other users perceive the interaction with chatbots as not convincing or engaging enough (Mimoun, Poncin, & Garnier, 2012). As chatbots run into problems and raise dissatisfaction in the customers, many chatbot services have already been discontinued (Brandtzaeg & Følstad, 2018; Gnewuch et al., 2017).

Characteristics of chatbots

As already mentioned before, trust plays an important role in the interaction between humans and chatbots (Corritore et al., 2003). Trust can be defined and determined by different factors, such as trust in the abilities of the chatbot or trust in privacy and safety of use of the chatbot (Przegalinska, Ciechanowksi, Stroz, Gloor, & Mazurek, 2019). A way to increase trust is

through the attribution of human behaviour or characteristics to the chatbot, in other words, anthropomorphisation (Qui & Benbasat, 2009). An example of this is the implementation of an avatar, which was found to increase trust in the chatbot (Angga, Fachri, Elevanita, Suryadi, & Agushinta, 2015). Some chatbots are designed to appear more humanlike by imitating human behaviour while exchanging texts with the users by delaying the time they take to respond to a message (Gnuwech, Morana, Adam, & Maedche, 2018). In that, they try to simulate the time it would take a person to formulate a response. This not only makes the chatbot appear more humanlike, but it also cues a reaction based on social expectations, leading to more satisfaction within the user (Gnuwech et al., 2018). Making chatbots more humanlike also increases the perceived social presence in the user, which is the perception that one is genuinely communicating with a medium such as a chatbot because it appears sociable, warm, sensitive, personal, or intimate (Qui & Benbasat, 2009). The perception of social presence can be effectively achieved by using speech as a means of communication (Qui & Benbasat, 2009). In many cases, a speech-based interface is not possible, and a text-based interface is used instead. Nevertheless, the use of natural language, as is the case for a text-based interface, still is beneficial in that it aids handiness and makes interactions less complicated (Gnuwech, Moran, & Maedche, 2018). Another antecedent for the perception of social presence was found in high message interactivity, which in turn also led to higher satisfaction ratings after the use of chatbots (Go & Sundar, 2019). Message interactivity can be defined by the degree of the contingency of messages upon the previous message and the other messages before that (Rafaeli, 1988). In other words, higher message interactivity makes a conversation feel more ongoing and interactive because responses are related to the messages that were exchanged beforehand. While it is possible to build trust in the user by making the system more humanlike, a more important factor in establishing trust is seen in having had prior experiences with said system (Gefen, 2000). Because trusting in a systems' capability to handle a certain problem is dependent on the context the system is used in, familiarity with the system constitutes the foundation that trusts eventually builds on.

Familiarity with the use of chatbots

The more familiar users become with chatbots, the more their trust in that system increases (Gefen, 2000). Generally, familiarity can be defined as knowledge of what, why, where, and when others do what they do (Gefen, 2000). Through the accumulation of experiences with a chatbot, users become more familiar with the technology (Mimoun, Poncin, & Garnier, 2017). Not only is familiarity with a chatbot a precondition for trust, but it also encapsulates the users understanding of how to use the chatbot. Thus, having prior experiences in the use of chatbots will make all future interactions easier because the user already knows about the functions a chatbot may have and how to access those functions. As familiar users know how to use a chatbot, using them will more likely lead to the desired outcome and therefore a more satisfactory experience. Prior experiences are of high importance in this case, as it supplies the user with first-hand information about the functionality, which is commonly considered to be more reliable in comparison to information that was gained indirectly (McKnight, Cummings, & Chervany, 1998). Also, the attitude towards the technology that is formed while directly experiencing it is more readily accessible (McKnight et al., 1998). More familiarity leads to more knowledge about the technology and how it operates and thus facilitates the interaction with it (Mimoun et al., 2017). In addition, familiarity also influences the evaluation of the technology. Familiarity is not only defined by the experiences one has made in the use of chatbots, but it can also occur in the form of knowledge or expertise about the artefact (Mimoun et al., 2017). This distinction is emphasised by Brucks (1985), who argues that similar experiences can teach different people different things, wherefore their resulting behaviour will also be different. Hence, knowledge defined by the experiences through which they have been gained is not as good a predictor of behaviour as the knowledge a person has about the functionality of the artefact (Brucks, 1985). In the case of chatbots, people with higher skills in the use of the internet need less time and effort to use a chatbot and rate chatbots as more useful (Mimoun et al., 2017). The fast and effortless handling of chatbots by people with more knowledge can be explained by the relevant information that is readily available to them (Alba, 1983). As an example of this, users with a high amount of prior knowledge may find it easier to evaluate the responses they get for the questions they pose, whereby using that information has a lower

cognitive cost (Brucks, 1985). Especially in complex situations, prior knowledge helps the user to interact with the chatbot in a more effective way (Brucks, 1985). This is supported by the notion that to know what question to ask, one first needs to know what is not known (Miyake & Norman, 1979). A mediating variable in asserting better skills to people with more familiarity with and knowledge about chatbots may be the involvement or interest these people have in chatbots (Brucks, 1985). Perceiving oneself to be knowledgeable about the use of chatbots may furthermore also result in more self-confidence when using a chatbot, thus also contributing to better performance (Brucks, 1985).

Quality of use

While identifying what qualities make a user more capable of using a chatbot is important, it is more helpful to identify what qualities a chatbot should have so that any user can easily use it. Moreover, a chatbot should not only be easy to use but also be useful in that it helps the user to achieve their intended goal (Bevan, 1995). These qualities are measured as usability which is defined as the "extend to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO, 2018). As the definition states, usability is always determined by the context in which it is used as it determines how appropriate a system, product or service is to this specific context. There are, however, measurements of usability that can be used in a reliable manner across different contexts and interfaces by standardised measures of usability. One of the most popular scales to assess perceived usability in a quick and reliable manner is the System Usability Scale (SUS) which has 10 items measured on a 5-point Likert scale, half of which have a negative tone and the other half have a positive tone (Brooke, 1996). The Computer System Usability Questionnaire (CSUQ) is another popular scale for perceived usability with 16 items with three subscales (System Usefulness, Information Quality, and Interface Quality) measured on a 7-point Likert scale (Lewis, 2002). Both scales demonstrated excellent reliability with Cronbach's α of 0.97 and 0.93 respectively (Lewis, 2018). The shortest scales for perceived usability are the 4-item usability metric for user experience (UMUX; Bosley, 2013) and the shortened two-item version of the usability metric for user experience (UMUX-Lite; Lewis, Utesch, & Maher, 2013). While the 4 items

of the UMUX are mixed in tone, the UMUX-Lite only uses the positive-tone items. The first of those two items asks the user about the quality of the system's functions ("The system's capabilities meet my requirements"), while the second item assesses the ease of use of the system ("The system is easy to use"). The two items of the UMUX-Lite correspond to the constructs of the Technology Acceptance Model (TAM; Davis, 1989) from the field of market research which is used to assess usefulness and ease of use of systems, which influences the likelihood of future use (Lewis, 2018). Research on the psychometric properties of the UMUX-Lite demonstrated acceptable reliability with estimates of Cronbach's α between 0.77 and 0.86 (Berkman & Karahoca, 2016; Borsci et al., 2015; Lewis et al., 2013, 2015). The concurrent validity and sensitivity of the UMUX-Lite were also shown to be acceptable. Despite the ability of these scales to assess usability, to determine what constitutes the usability of a chatbot, the users, their goals, and the context in which chatbots are used must also be considered. Such factors were determined through a systematic literature review by Tariverdiyeva & Borsci (2019) and translated to an initial scale to assess user satisfaction for chatbots. After consolidation with experts and end-users on the importance of the specific features. Balaji & Borsci (2019) developed the user satisfaction questionnaire (USQ) with 42 items that measures the satisfaction of users after an interaction with a chatbot. Later, the USQ was shortened to 15 items with 5 underlying factors as the Chatbot Satisfaction Scale (Appendix A), so it would be less of a burden for respondents (Borsci et al., under review).

The present study

The present study aims to confirm the current factorial model of the Chatbot Satisfaction Scale. To do so, the internal and external validity of the scale needs to be evaluated. To determine the internal validity of the Chatbot Satisfaction Scale its factorial structure will be evaluated by performing confirmatory factor analysis. Additionally, it will be compared to the UMUX-Lite (Lewis, Utesch, & Maher, 2013) to verify the external validity of the scale. The first two research questions will thus be:

RQ1 - Can the current factorial structure established by exploratory factor analysis on the Chatbot Satisfaction Scale be verified?

RQ2 - Do the results of the Chatbot Satisfaction Scale correlate with the results of the UMUX-Lite?

Moreover, the Chatbot Satisfaction Scale was translated to German so that it could be used by a broader range of researchers. To validate if the translation is correct and more importantly also retained the same psychometric properties, the second aim of this study is to compare the English version of the Chatbot Satisfaction Scale to the German version. Hence, the third research question is:

RQ3 - Do the results of the German translation of the Chatbot Satisfaction Scale correlate to the results of the original English version?

As already demonstrated earlier, familiarity and increased knowledge about the use of chatbots may lead to a more satisfactory interaction with the system in use. To test this hypothesis, the fourth research question is:

RQ4 - Do people with more familiarity with the use of chatbots give higher satisfaction ratings compared to people with less familiarity?

Methods

Participants

Using the BMS Test Subject Pool system SONA and convenience sampling 74 volunteers were recruited to participate in the experiment (*M*age = 29.21, *SD*age = 13.94). Psychology and Communication Science Students from the University of Twente who signed up using SONA were able to earn credits through their participation. 46 of the participants experimented under supervision, while the other 28 participants were not supervised during participation. Due to incomplete responses, the data of 18 participants were omitted for the analysis. Of the remaining 56 participants, 35 (62.5%) were female and 21 (37.5%) were male. 39 participants were German, 2 were Dutch and the remaining 15 had other nationalities (4 Columbian, 2 Italian, 2 American, 1 Vietnamese, 1 Romanian, 1 Salvadorian, 1 Peruvian, 1 n.d). Approximately 11% of the participants were either extremely or very familiar with chatbots, while 32% were moderately familiar, 43% slightly familiar and 14% not familiar at all. 71% of the participants said that they had definitely or probably used a chatbot before, 14% were unsure and 15% had not definitely or probably never

used a chatbot. 13% reported that they never use chatbots, 71% reported occasional use of chatbots and 16% used chatbots on a daily to weekly basis. Lastly, participants were asked to indicate how confident they felt using a chatbot, whereas 61% responded to be very or moderately confident, 28% slightly confident and 11% not confident at all. As each of the participants interacted with 10 different chatbots and assessed their usability a total of 560 observations was collected. The experiment received ethical consent from the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente.

Materials

The meetings with the participants were held online within Zoom (n.d.). For the data gathering, an online questionnaire designed within Qualtrics (n.d.) was used. The participants were able to access the questionnaire with a link that took them to Qualtrics (n.d.). There they were presented with the shortened 15-item version of the Chatbot Satisfaction Scale (Appendix B) developed by Borsci et al. (under review) and the UMUX-Lite after each interaction with a chatbot. While the Chatbot Satisfaction Scale uses a 5-point Likert scale, the UMUX-Lite uses a 7-point Likert scale. Both questionnaires were translated into German (Appendix C & Appendix D). Also presented in Qualtrics (n.d.) were the descriptions of the tasks as well as the links to the websites where the chatbots could be found (Appendix E). Many of the tasks were similar to the tasks that were used by van den Bos and Borsci (2021) and were additionally translated into German and Spanish.

Tasks

Participants received the task to locate and interact with 10 different chatbots and subsequently fill out two scales about their experience i.e., the Chatbot Satisfaction Scale and the UMUX-Lite. To begin with, participants had to read a scenario that informed them about the goal of their interaction with one of the chatbots. Additionally, they were given a link that they had to copy and paste in the address bar of their browser after reading the scenario. Opening the link took the participants to the website where the chatbot could be found. They had to locate the chatbot by themselves and start a conversation with it. The interaction with each chatbot entailed asking the chatbot questions until the objective of the associated scenario was reached. Whether the objective

was reached was or not determined by the participant. After each objective was reached, the participant had to navigate back to the survey within Qualtrics and start filling out the scales. The first scale presented was the UMUX-Lite and afterwards the 15 items of the Chatbot Satisfaction scale in a randomised order.

Procedure

The experiment was initially carried out under supervision. After the response rate was not as high as expected, the experiment was offered to be filled out without supervision, giving the participants more flexibility when choosing a time to participate. The supervised participants were sent a link to the Zoom (n.d.) meeting, fifteen minutes before the scheduled start of the online session. As soon as the participants joined the meeting, the researcher provided them with a link that took them to the survey within Qualtrics (n.d.). Unsupervised participants were provided with a link and were able to access the survey at a time of their preference. Participants were informed that they could choose between an English and a German version of the survey and what the goal of having two different versions is. Then, the participants could choose one of the three languages and go on to the survey. As the first part of the survey, the participants had to read the informed consent form (Appendix F) and either actively give consent by clicking yes or decline. In the case that a participant declined the informed consent, that concluded the session. If the participants actively gave their consent, they moved on to questions about their demographics and subsequently to a questionnaire about their prior experiences and familiarity with chatbots (Appendix G). Before starting to interact with the chatbots, the researcher explained the tasks to the supervised participant and gave them time to ask any questions they had. Unsupervised participants were informed about the tasks in the form of a text. It was emphasised during the explanation, that the aim of the scales is not to assess the participants' performance during the interaction with the chatbots but only the satisfaction level after the interaction. The participants continued with the ten different tasks, interacting with the chatbots, and filling out the two scales after each interaction with a chatbot. The researcher stood by during that period to offer the supervised participant the opportunity to ask questions and to clear up any uncertainties that could come up. Upon completion of all the tasks, the supervised participants were again asked if they had any

questions, informed about ways to contact the researcher after the session and thanked for their participation. Again, the unsupervised participants were informed by text on how to contact the researcher.

Data Analysis

The data was transferred from Qualtrics (n.d.) to Excel as comma-separated values. There, all rudimentary data was excluded, items were labelled, and the dataset was rearranged to be compatible with R (v4.0.5; R Core Team, 2021). Two incomplete lines of data were retained in the dataset as only responses for one of the ten chatbots were missing. Afterwards, the dataset contained a total of 558 lines of data. The dataset included the data from the English version as well as the data generated from the German and Spanish (Kerwien-Lopez & Borsci, 2021) version of the scale.

To answer the first research questions and thus assess whether the factor structure of the Chatbot Satisfaction Scale is comparable to the one found in the previous study by Borsci et al. (under review) confirmatory factor analysis was performed using the Lavaan package (Rosseel, 2012). First, the data were tested on normality graphically by creating a density plot and a Q-Q plot using the dplyr package (Wickham, 2018) and the ggpubr package (Kassambara, 2020). A statistical test for normality was conducted with a Shapiro-Wilk test showing that the data was not normally distributed. Consequently, the data analysis was continued using non-parametric statistics. Thus, a Mann-Whitney U test was performed to test whether there was a significant difference in responses from the supervised and unsupervised participants. To test the internal consistency of the Chatbot Satisfaction Scale, Cronbach's alpha was computed using the Psych package (Revelle, 2020). Eventually, the confirmatory factor analysis was performed. To determine how good the fit of the model is, the fit indices were looked at. More specifically, the comparative fit index (CFI), the absolute fit index (RMSEA), the absolute measure of fit (SRMR), the Tucker-Lewis index (TLI), the Expected Cross-validation index (ECVI), the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and Chi-squared were investigated. For a good model fit, RMSEA is preferably below 0.06, SRMR below 0.07, CFI and TLI higher than 0.95.

When comparing between models, lower scores on ECVI, AIC and BIC indicate a better model fit. Also, the chi-square statistic should be non-significant.

To test the concurrent validity of the 15-item version of the Chatbot Satisfaction Scale, the second research question aimed at comparing the results of the Chatbot Satisfaction Scale with the results of the UMUX-Lite. Before this could be done, the reliability of the translation of the UMUX-Lite was assessed by computing Cronbach's alpha. To see whether there was a significant difference between the two scales, a t-test was performed. Afterwards, mean scores were calculated for each line of data from the Chatbot Satisfaction Scale and the UMUX-Lite. Kendall's rank correlation coefficient was then used to calculate the correlation with the MASS package (Venables, 2002).

The third research question was aimed at comparing the German translation of the Chatbot Satisfaction Scale to the original English version. Firstly, Cronbach's Alpha was calculated for each version to check the intrinsic validity for each translation of the Chatbot Satisfaction Scale. Then, a t-test was used to test for a significant correlation between the two translations. Also, correlations were calculated using Kendall's rank order correlation.

Lastly, the relationship between familiarity and satisfaction was analysed. First, the reliability of the survey assessing familiarity was tested with Cronbach's alpha. To test the hypothesis that more familiarity leads to higher satisfaction ratings, summarised mean scores were calculated for all the responses of each participant from the Chatbot Satisfaction Scale. Additionally, familiarity scores were calculated for each participant based on their responses to the survey. Subsequently, Kendall's rank order correlation was used to compare familiarity with satisfaction scores. A linear regression model was used to test the hypothesis that different levels of familiarity with chatbots affect participants' satisfaction ratings after the interaction experience.

Results

This section will be divided into analysis on the Chatbot Satisfaction Scale, a comparison between the Chatbot Satisfaction Scale and the UMUX-Lite, analysis on the German translation of the Chatbot Satisfaction Scale, and finally analysis on the effect of familiarity on the Chatbot Satisfaction Scale.

Chatbot Satisfaction Scale

Test for normality

The graphical test for normality was performed on all the responses for the Chatbot Satisfaction Scale. The density plot (Figure 1) and the Q-Q plot (Figure 2) indicated that the data was not normally distributed. This observation was confirmed by the output of the Shapiro-Wilk test as the p-value was smaller than 0.05, wherefore the null-hypothesis that the data is normally distributed was rejected.

Figure 1

Density plot to test for normal distribution of the data.

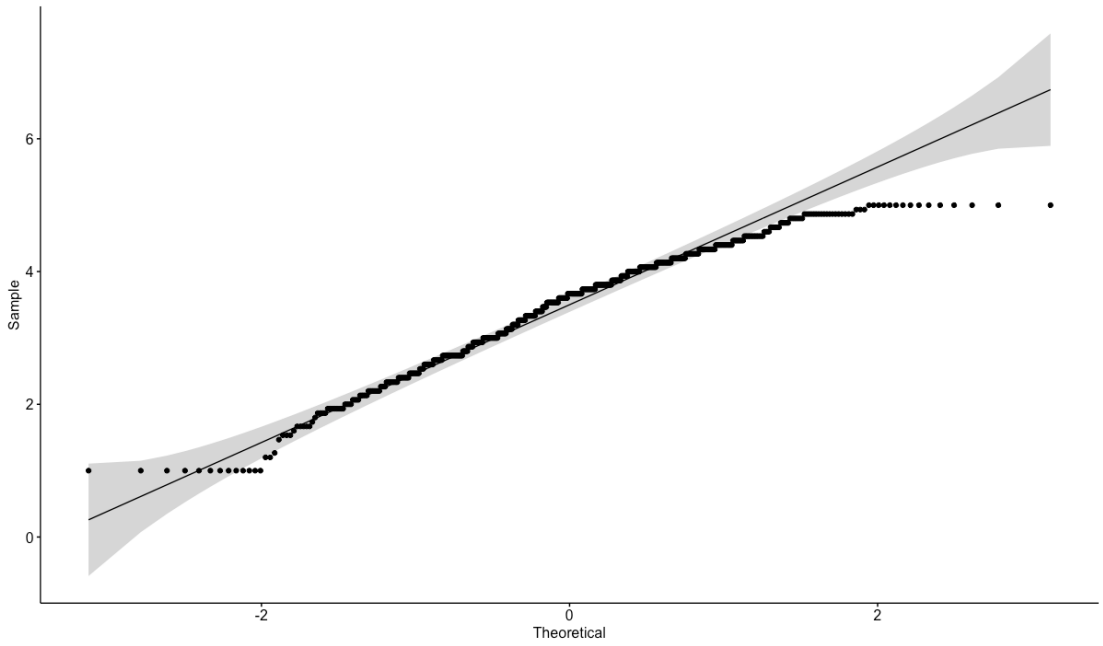
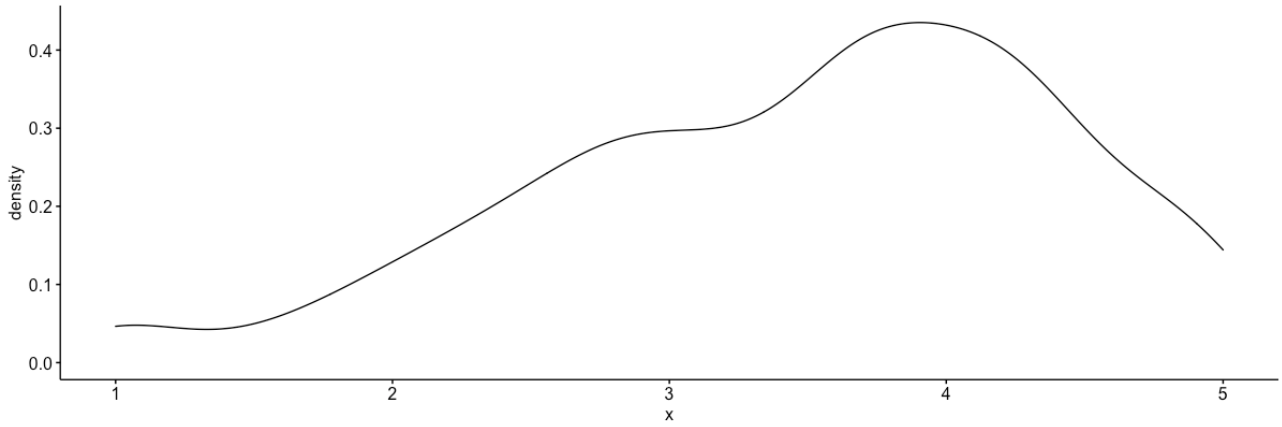


Figure 2

Q-Q plot to test for normal distribution of the data.



Manipulation check for supervised and unsupervised participants

A Mann-Whitney U test indicated that the responses of supervised participants (*Mdn* = 4) were not significantly different from the responses of unsupervised participants (*Mdn* = 4), *W* = 33806, *p* = .534.

Confirmatory factor analysis

The reliability of the Chatbot Satisfaction Scale was shown to be high with a Cronbach's α = 0.93. Factor 1 "Perceived accessibility to chatbot functions" assessing how easy it was for participants to become aware of the chatbot function and how to access it had a Cronbach's α = 0.89. The second factor "Perceived quality of chatbot functions" which measures conversation flow in regard to clarity and keeping track of context, but also how well the chatbot informs the user of its capabilities and makes references to other websites or services had a Cronbach's α = 0.9, The highest reliability was found in factor 3 which evaluates the chatbot's competence to understand what the user wants and give the right the amount of appropriate of information as a response with Cronbach's α = 0.93. Factors 4 and 5, which assessed the chatbot's capability to inform the user of privacy issues and respond in a timely manner respectively, both only had one item. In comparison, the overall reliability that was tested by Borsci et al. (under review) for this scale is Cronbach's α = 0.87. The factor loadings from the confirmatory factor analysis on the model

proposed by Borsci et al. (under review) can be found in Table 1. To improve on the initial model, it was modified based on the factor loadings of the items, r-squared and modification indices.

Table 1

Standardised factor loadings for all the items for the initial model with a 5-factor structure.

Item	Factor 1 "Perceived accessibility to chatbot functions"	Factor 2 "Perceived quality of chatbot functions"	Factor 3 "Perceived quality of conversation and information provided"	Factor 4 "Perceived privacy and security"	Factor 5 "Time response"
Item 1	.951				
Item 2	.838				
Item 3		.895			
Item 4		.711			
Item 5		.653			
Item 6		.807			
Item 7		.665			
Item 8		.680			
Item 9		.778			
Item 10			.898		
Item 11			.901		
Item 12			.838		
Item 13			.888		
Item 14				1	
Item 15					1
Median	4	4	4	2	4

All the items for the initial model had high factors loadings. Only items 5, 6, and 8 had a factor loading below 0.7. Also, the model fit indices for the initial model were already good, with only chi-square being significant and RMSEA being too high. The fit indications for the first model with 15-items can be found in Table 2.

Table 2

Model fit indications for the initial model with 15-items from Borsci et al. (under review)

Model	CFI	RMSEA	SRMR	TLI	ECVI	AIC	BIC	X2 (df)
1	.952	.080	.040	.938	.812	22042.44	22206.77	(82) 377.37

Note. CFI = Comparative fit index. RMSEA = Absolute fit index. SRMR = Absolute measure of fit. TLI = Tucker-Lewis index. ECVI = Expected cross-validation index. AIC = Akaike information criterion. BIC = Bayesian information criterion. X2 = Chi-square.

The modification indications showed that there was a covariation between the items 6 and 8, 5 and 6, as well as 5 and 8. Looking at the items this makes sense, as all these items seem to be related to the flow of the conversation and how well the chatbot could keep track of previously given information. Adding a covariation between these items to further specify the model indicated that chi-square could be lowered. Therefore, a new model was created in which covariations between items 5 ("The interaction with the chatbot felt like an ongoing conversation"), 6 ("The chatbot was able to keep track of context"), and 8 ("The chatbot could handle situations in which the line of conversation was not clear") were specified (see Table 3). The fit indications for this second model with 15-items and specified covariances between items 5, 6, and 8 can be found in Table 4. This model already indicated a slightly better fit but did not improve chi-square and RMSEA to a satisfactory level.

Table 3

Overview of all the models used in the confirmatory factor analysis

Model	Factors and corresponding items						Specified covariances
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	
1	Q1, Q2	Q3, Q4, Q5, Q6, Q7, Q8, Q9	Q10, Q11, Q12, Q13	Q14	Q15		

2	Q1, Q2	Q3, Q4, Q5, Q6, Q7, Q8, Q9	Q10, Q11, Q12, Q13	Q14	Q15	Q5, Q6, Q8
3	Q1, Q2	Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13		Q14	Q15	
4	Q1, Q2	Q3, Q4, Q6, Q7, Q9	Q10, Q11, Q12, Q13	Q14	Q15	
5	Q1, Q2	Q3, Q6, Q9	Q4, Q7	Q10, Q11, Q12, Q13	Q14	Q15

Table 4

Model fit indications for the modified model with 15-items and specified covariances between items 5, 6, and 8.

Model	CFI	RMSEA	SRMR	TLI	ECVI	AIC	BIC	X2 (df)
2	.967	.067	.036	.957	.647	21950.11	22127.41	(79) 279.03

Note. CFI = Comparative fit index. RMSEA = Absolute fit index. SRMR = Absolute measure of fit. TLI = Tucker-Lewis index. ECVI = Expected cross-validation index. AIC = Akaike information criterion. BIC = Bayesian information criterion. X2 = Chi-square.

To further improve the model fit a comparable factor structure to the one used by Silderhuis & Borsci (2020) was applied. In their model factor 2 “Perceived quality of chatbot functions” and factor 3 “Perceived quality of conversation and information provided” only made up one factor assessing communication quality. This was also supported by the observation that items of factor 2 covariates to factor 3 and vice versa. Thus, a third model was created in which all the items for factor 2 and factor 3 were associated with only one factor (see Table 3). The fit indications for this

third model in which all items for factors 2 and 3 were associated in one factor can be found in Table 5. Although this resulted in a better fit compared to the initial model from Borsci et al. (under review), there was no improvement in comparison to the previous model with 5 factors where covariances for items 5, 6, and 8 were specified. No significant difference could be observed in the factor loadings. Ultimately this third model with 4 factors was rejected because it did not show a better fit compared to the second model with 5 factors.

Table 5

Model fit indications for the modified model in which the items from factors 2 and 3 are associated in one factor assessing communication quality

Model	CFI	RMSEA	SRMR	TLI	ECVI	AIC	BIC	X2 (df)
3	.953	.079	.043	.941	.796	22033.01	22193.01	(83) 369.94

Note. CFI = Comparative fit index. RMSEA = Absolute fit index. SRMR = Absolute measure of fit. TLI = Tucker-Lewis index. ECVI = Expected cross-validation index. AIC = Akaike information criterion. BIC = Bayesian information criterion. X2 = Chi-square.

Another attempt to improve the model fit constituted extracting items 5 ("The interaction with the chatbot felt like an ongoing conversation") and 8 ("The chatbot could handle situations in which the line of conversation was not clear") as those showing the lowest factor loadings (see Table 1). As already mentioned before items 5 and 8, together with item 6 ("The chatbot was able to keep track of context"), all measure how well the chatbot can use previously used and given information to ensure a good flow of conversation. Because item 6 had the highest factor loading between the three items, it appeared that it could best explain what these three items are assessing. Hence, a fourth model was created in which items 5 and 8 were extracted from the model (see Table 3). The model fit indications for this model with 13 items can be seen in Table 6. This fourth model in which items 5 and 8 were extracted showed improvement in fit in comparison to all other models that were tested before. While Chi-square was lower for this fourth model in comparison to all previously tested models it was still not non-significant. RMSEA for model 4 was the same as for model 2 in which covariances for items 5, 6, and 8 were added. Due to the significant chi-square

statistic an RMSEA the fourth model with 13 items still did not meet the satisfactory criteria for a good model fit.

Table 6

Model fit indications for the fourth model with 13 items.

Model	CFI	RMSEA	SRMR	TLI	ECVI	AIC	BIC	X2 (df)
4	.973	.067	.031	.964	.482	18917.64	19064.67	(57) 200.737

Note. CFI = Comparative fit index. RMSEA = Absolute fit index. SRMR = Absolute measure of fit. TLI = Tucker-Lewis index. ECVI = Expected cross-validation index. AIC = Akaike information criterion. BIC = Bayesian information criterion. X2 = Chi-square.

By looking at the explained variance of the items from the model with 13 items it became apparent that item 4 ("I was immediately made aware of what information the chatbot can give me") and item 7 ("The chatbot was able to make references to the website or service when appropriate") were not sufficiently explained by the model. After inspecting the items, it was determined that either items 4 and 7 or the remaining items from the factor "Perceived quality of chatbot functions" (items 3, 6, and 9) could be better explained by another factor. Thus, another model with six factors was created in which items 4 and 7 were explained by a factor other than the one items 3, 6, and 9 were explained by (see Table 3). Apart from a higher BIC and a steady RMSEA and TLI, this fifth model showed minor improvements in comparison to the previously tested model. Despite a lower chi-square statistic in comparison to the previous model, also the fifth model with six factors does not meet all satisfactory criteria for a good model fit. However, especially ECVI and AIC were significantly lower in comparison to the other models tested before in this factor analysis and also slightly lower in comparison to the previous model, indicating that this model is the best fit. In Table 7 the model fit indications for the model with 13 items and 6 factors are presented and the factor loadings are shown in Table 8. A graphical representation of the factor structure is presented in Figure 3. For this new model, the overall reliability is to be considered excellent, Cronbach's α = 0.93.

Table 7

Model fit indications for the fifth model with 13 items and six factors

Model	CFI	RMSEA	SRMR	TLI	ECVI	AIC	BIC	X2 (df)
5	.976	.067	.029	.964	.464	18907.82	19076.47	(52) 180.909

Note. CFI = Comparative fit index. RMSEA = Absolute fit index. SRMR = Absolute measure of fit. TLI = Tucker-Lewis index. ECVI = Expected cross-validation index. AIC = Akaike information criterion. BIC = Bayesian information criterion. X2 = Chi-square.

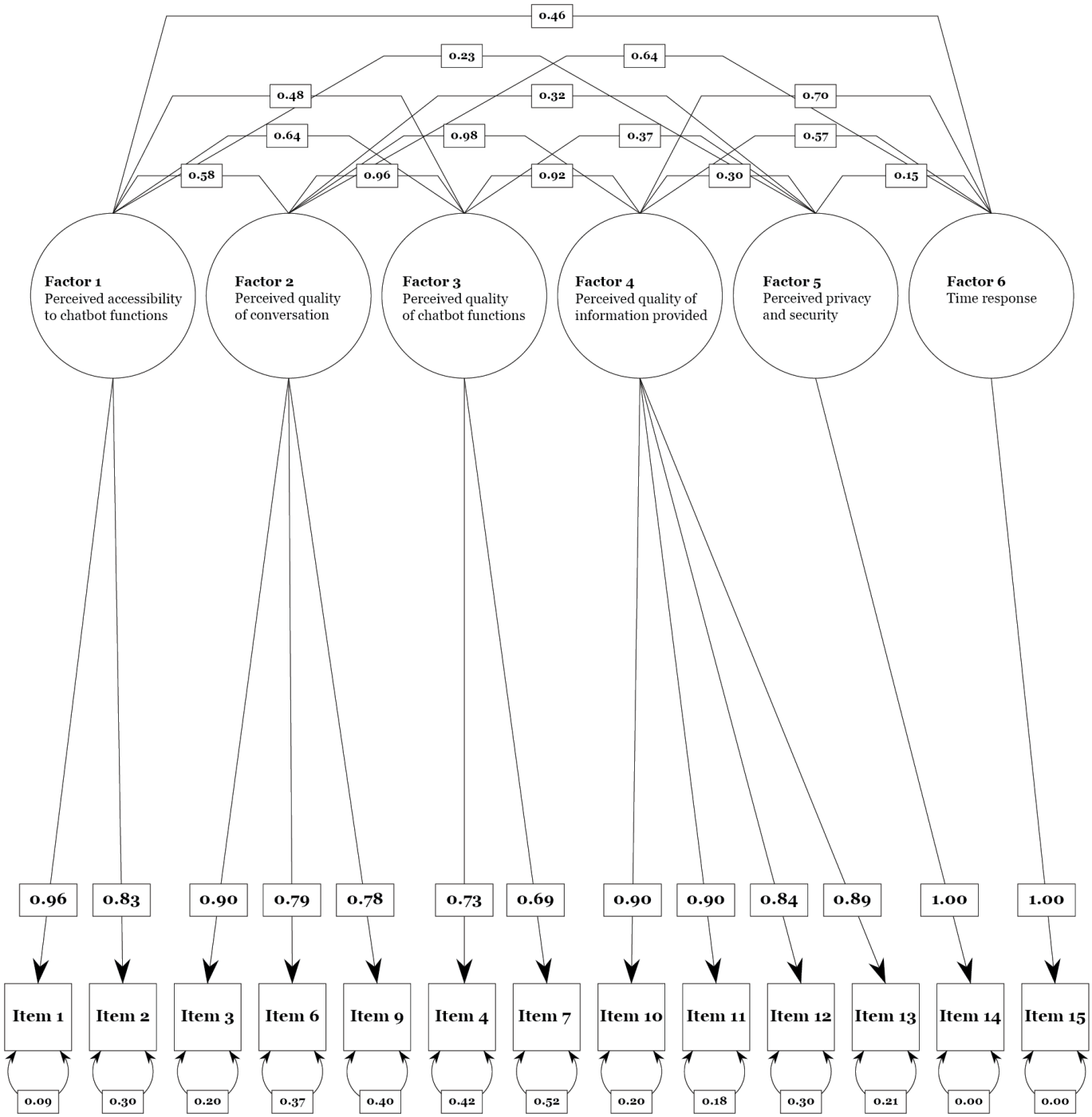
Table 8

Standardised factor loadings for the sixth model with 13 items.

Item	Factor 1 "Perceived accessibility to chatbot functions"	Factor 2 "Perceived quality of conversation"	Factor 3 "Perceived quality of chatbot functions"	Factor 4 "Perceived quality of information provided"	Factor 5 "Perceived privacy and security"	Factor 6 "Time response"
Item 1	.956					
Item 2	.834					
Item 3		.897				
Item 4			.729			
Item 6		.791				
Item 7			.692			
Item 9		.777				
Item 10				.895		
Item 11				.903		
Item 12				.839		
Item 13				.890		
Item 14					1	
Item 15						1
Median	4	4	4	4	2	4

Figure 3

Model 5 with standardised factor loadings and covariances between the factors.

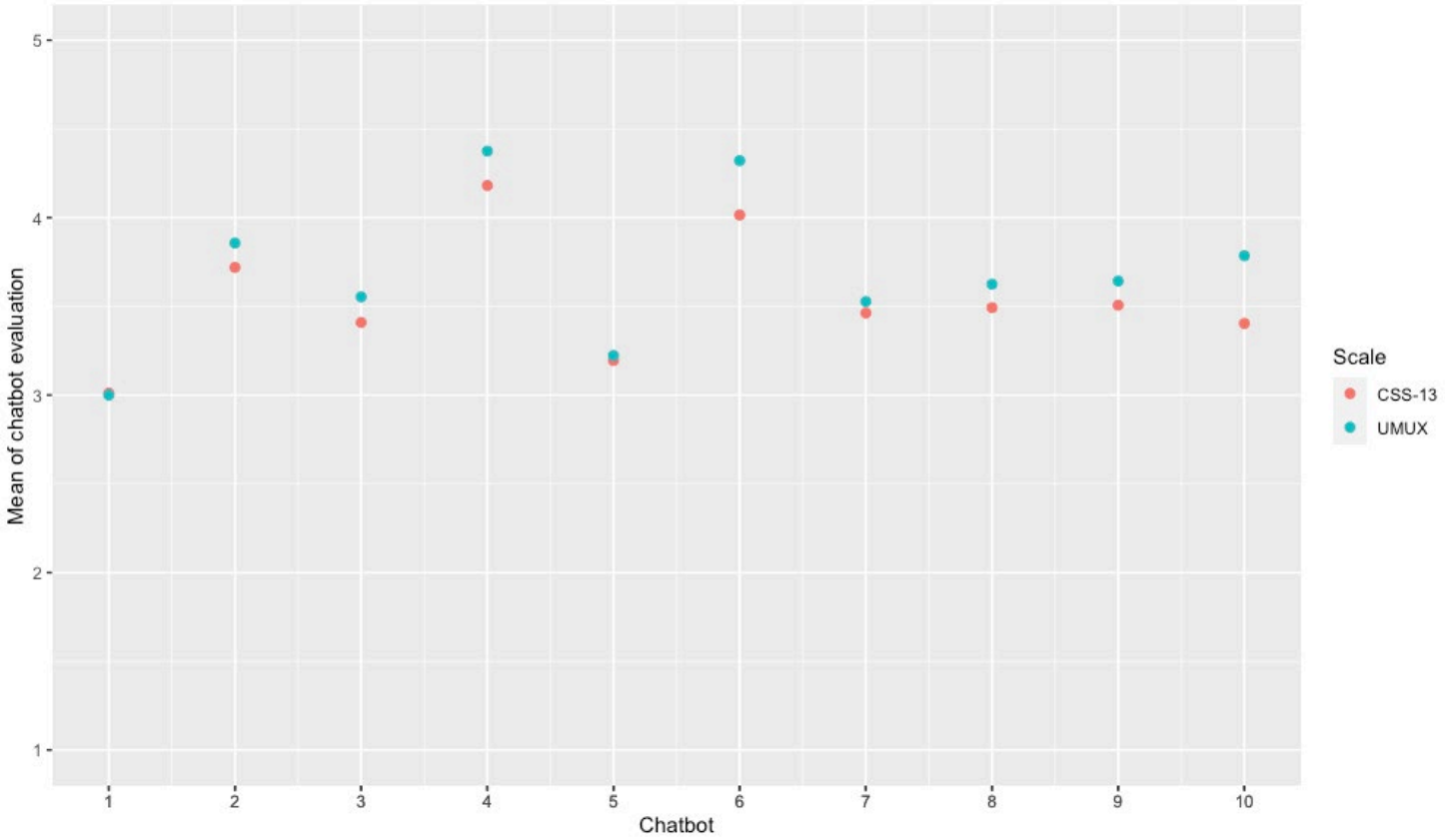


Comparing the Chatbot Satisfaction Scale with 13 items and the UMUX-Lite

The overall reliability of the German translation of the Chatbot Satisfaction Scale with 13 items (CSS-13) was good (Cronbach's α = 0.95) and quite in line with the original English version (Cronbach's α = 0.92). When comparing the overall score of the CSS-13 with the results of the UMUX-Lite with a Mann-Whitney U test no significant difference between the two scales was found, $W = 148752$, $p = 0.19$. Kendall's rank order analysis suggests that there is a positive correlation between the scales, $r_\tau = 0.78$, $p < 0.01$. The means of the responses for both scales in graphically displayed in Figure 4.

Figure 4

Plot displaying the results of a Compare Group Means analysis on the mean satisfaction scores of the UMUX-Lite and the Chatbot Satisfaction Scale with 13 items.



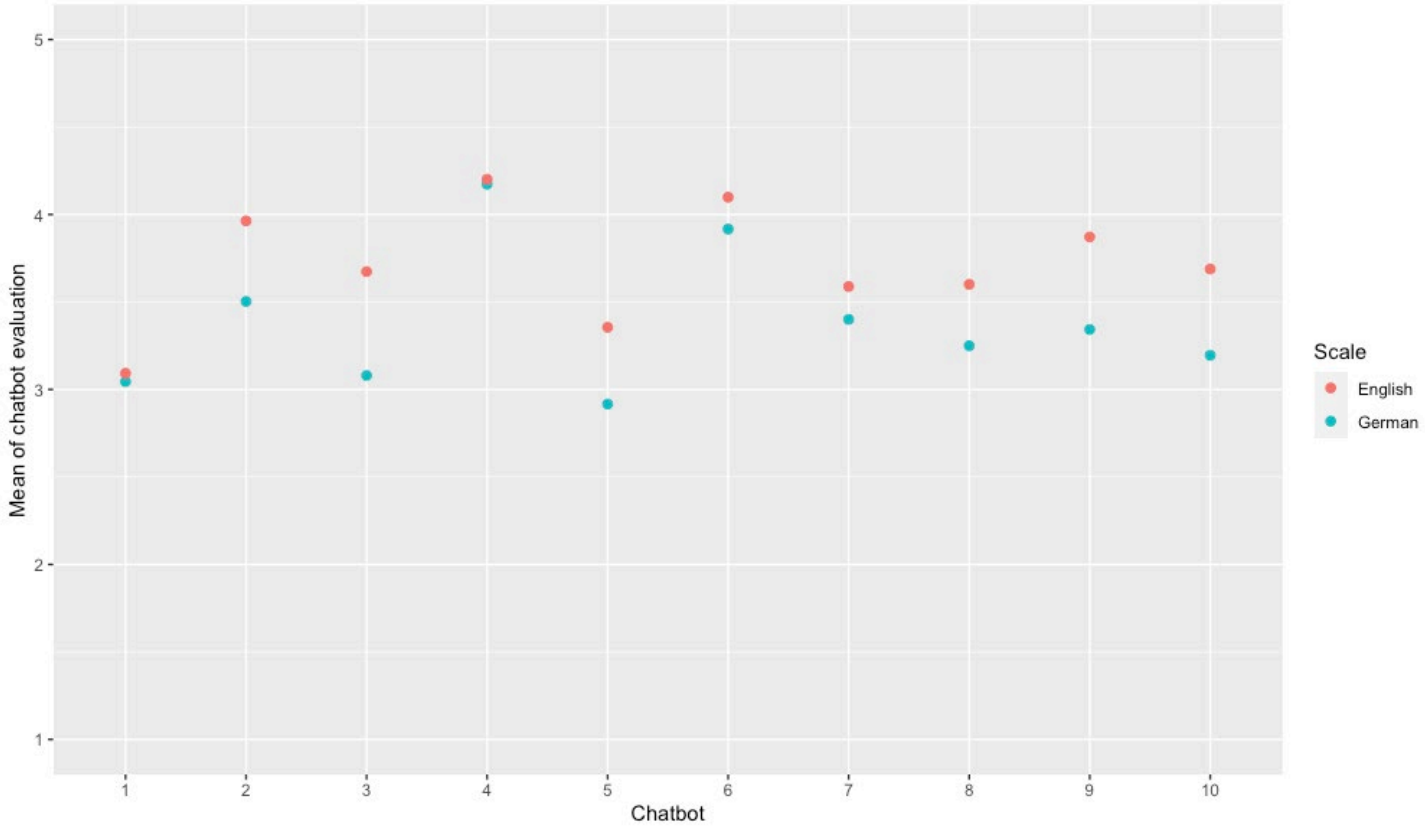
Note. "CSS-13" is used as an indicator for the scores on the Chatbot Satisfaction Scale with 13 items and "UMUX" as indicator for the scores on the UMUX-Lite.

Comparing English and German version of the CSS-13

The reliability of the German and the English translation of the CSS-13 was found to be excellent, with Cronbach's α = 0.95 and α = 0.92 respectively. A Mann-Whitney U test showed that there was a significant difference between the two translations of the scale, $W = 20280$, $p < 0.01$. A positive correlation between the two versions was shown during Kendall's rank order correlation analysis, $r_\tau = 0.69$, $p < 0.01$. A graphical presentation of the mean scores of each translation of the CSS-13 can be found in Figure 5.

Figure 5

Plot displaying the results of a Compare Group Means analysis on the mean satisfaction scores of both translations of the Chatbot Satisfaction Scale with 13 items.



Effect of familiarity on the CSS-13

The test of internal consistency for the questionnaire assessing familiarity showed overall good reliability, Cronbach's α = 0.81. A Kendall's rank order correlation analysis was performed by

computing a mean satisfaction score for every participant with all the responses to the CSS-13 for all the 10 chatbots. The descriptive statistics for the dataset can be found in Table 10. The correlation analysis with the mean familiarity score showed no correlation to the satisfaction scores, $r_\tau = 0.068$, $p = 0.469$. None of the other combinations of items from the familiarity questionnaire showed a significant correlation with the scores on the CSS-13 either. The strongest but also non-significant correlation was found between the item assessing confidence in the use of chatbots or conversational agents, $r_\tau = 0.16$, $p = 0.119$. A linear regression model saw familiarity as a non-significant predictor for chatbot ratings on the CSS-13, $\beta = 0.02$, $p = 0.81$. A graphical representation of the linear regression with mean chatbot satisfaction ratings as a function of familiarity is presented in Figure 6. Also, the other factors that constituted the familiarity score, namely experience, $\beta = 0.004$, $p = 0.96$, and confidence, $\beta = 0.08$, $p = 0.26$, showed to be non-significant predictors for the chatbot satisfaction ratings. By examining individual scores of participants, it was observed that in many cases especially participants with little familiarity with chatbots and no prior experiences were more satisfied with the chatbots they used.

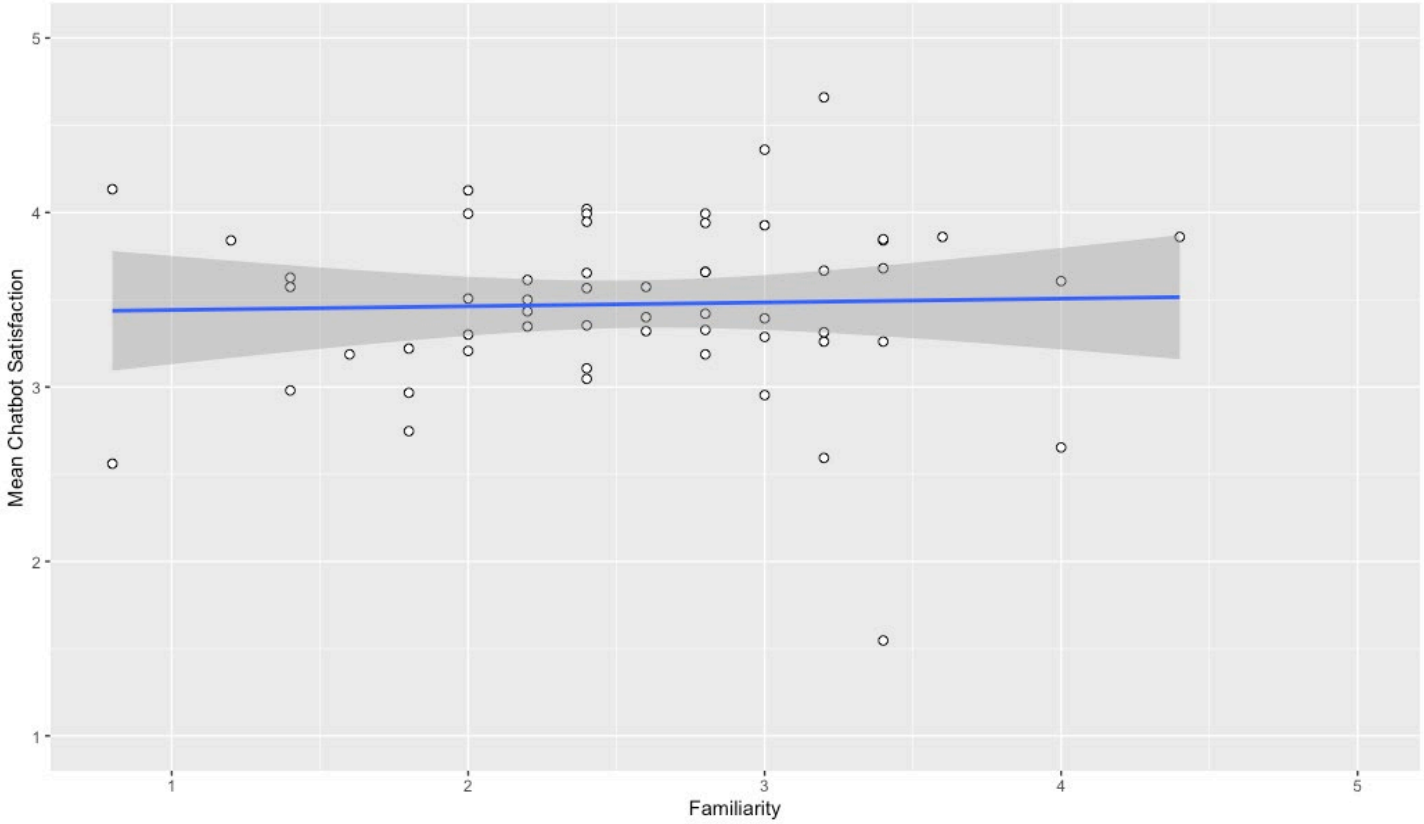
Table 10

Descriptive Statistics for mean satisfaction scores and familiarity ratings | N = 56

	Mean	SD	Min	Max
Familiarity rating	2.56	0.77	0.8	4.4
Mean Satisfaction Score	3.47	0.5	1.55	4.66

Figure 6

Mean chatbot satisfaction rating as a function of familiarity



Discussion

The aim of this study was to perform a confirmatory factor analysis of the psychometric properties of the Chatbot Satisfaction Scale (Borsci et al., under review) as a measure of perceived usability for chatbots. For this purpose, the scale was tested on how well it measures the implied concept. Furthermore, the study is aimed at concurrently validating the new scale by using the UMUX-Lite (Lewis, Utesch, & Maher, 2013). Moreover, to allow for more people to be able to use the scale, the CSS-13 was translated to German and a reliability analysis was performed. Last, it was investigated whether there is an effect of familiarity in the use of chatbots on satisfaction ratings.

Psychometric properties of the CSS-13

The first research question was: “Can the current factorial structure established by exploratory factor analysis on the Chatbot Satisfaction Scale be verified?” To answer this question a

confirmatory factor analysis was performed on the five-factor model proposed by Borsci et al. (under review). The results of the factor analysis on this initial model showed an already acceptable model fit, indicating that the Chatbot Satisfaction Scale in its original form measures the underlying concept of usability. Nevertheless, the confirmatory factor analysis demonstrated that the model could be improved by excluding two items and creating a new factor. The new 13-item Chatbot Satisfaction Scale will be especially useful in evaluating short interactions with chatbots in which only one task is to be completed. In situations where the interaction between user and chatbot is short, not many messages will be exchanged wherefore it is unlikely that the impression of an ongoing conversation will arise. Item 5 which assesses whether the interaction feels like an ongoing conversation and item 8 which assesses the ability of the chatbot to react appropriately at ambiguous times during the interaction appear to be designated for longer interactions in comparison to those tested within this study. However, both items measure the chatbots ability to incorporate previously given information, a quality that can contribute to perceived social presence within the chatbot (Go & Sundar, 2018). The ability is retained by item 6, which asks the user about the chatbots awareness of context in a more direct way. In short interactions with chatbots, this feels like a more appropriate approach as context can be established with the first message the user sends and does not require a longer interaction. The new underlying six factor structure of the CSS-13 will also allow for a more specific assessment of the individual features that constitute the usability of a chatbot. The items of factor 2 “Perceived quality of chatbot functions” are broken up into two new factors. First, the factor “Perceived quality of conversation” is created including the items 3, 6, and 9 which all assess how smoothly the conversation was progressing by measuring clarity, understandability, and the ability to integrate previous information of the chatbot. Second, items 4 and 7 were associated with the factor “Perceived quality of functions” because they relate more to the quality of the chatbot to aid the user in achieving their goal (Bevan, 1995). They do so in an indirect manner in that they do not provide the user with a direct answer to their question but rather present them with the tool to reach their goal. In contrast, the fourth factor “Perceived quality of information provided” assesses the more direct ability of the chatbot to provide the user with a solution to their problem. It does so by measuring how well the chatbot was able to translate the user’s prompt into an appropriate response. Moreover, the factor also assesses the quality of

the response in that it asks how much they trust the provided information and whether it was the appropriate amount. As demonstrated, the new 13-item model with an underlying 6 factor structure is more suited to a specific application on short interactions with chatbots and provides a more specific picture of the qualities that underlie the usability of a chatbot. To assess how satisfied users are with a system in an effective way, the context in which it is used should be specified to determine how appropriate it is to that context (Brooke, 1996). Thus, the more specified CSS-13 should help effectively in assessing usability in the context of short interactions with a chatbot and provide useful information to tailor chatbots to the needs of the user. Results also suggest, in line with the second research question ("Do the results of the Chatbot Satisfaction Scale correlate with the results of the UMUX-Lite?") that the CSS-13 strongly correlates with the UMUX-Lite. Overall, the mean satisfaction ratings measured with the UMUX-Lite were higher than the ones measured with the CSS-13. A reason for that may be that the UMUX-Lite measured the concept of usability in a more general way, whereas the CSS-13 is designed to assess the usability of chatbots and thus, considers more aspects of usability important for chatbots in the assessment.

Translation of the CSS-13 and familiarity

The third research question was: "Do the results of the German translation of the Chatbot Satisfaction Scale correlate to the results of the original English version?" There was a moderate strong correlation found between both translations of the scale. A compare group means analysis suggested a similar correlation. However, satisfaction scores tended to be lower for the people who used the German translation of the Chatbot Satisfaction Scale with 13 items. This indicates that the German translation fails to capture the way the original version was phrased and thus, should be revised. An item analysis may also be beneficial in finding out at what points the translation may have conveyed a different connotation towards the participant and thereby led to the difference in responses.

Finally, in line with the fourth research question ("Do people with more familiarity in the use of chatbots give higher satisfaction ratings compared to people with less familiarity?") mean familiarity scores were not found to predict the satisfaction ratings on chatbots. Taking the generally low scores of familiarity among the participants into account, it can be assumed that a

certain level of familiarity is necessary to evoke the hypothesised effect on satisfaction ratings. This assumption is also supported by the notion that familiarity is an underlying factor of trust in chatbots (Gefen, 2000). Users may have had ambiguous previous experiences with chatbots. While some may have built more trust through positive experiences with chatbots, others may have had more negative experiences which produced a disposition towards chatbots in them (Brucks. 1985). Rather than having familiarity in the form of having interacted with chatbots before, knowledge of how a chatbot works is often seen as a better indicator of familiarity with chatbots. As most participants reported to have little to no knowledge on how chatbots operate, this may be indicative of not finding an effect of familiarity on satisfaction. Although the mean familiarity score of the participants was not particularly high, participants gave high satisfaction ratings to the chatbot. This may be explained by the relatively young age of the participants, which lets to assume that most were quite skilled in the use of the internet. People who are experienced in the use of the internet often rate chatbots as more useful because they need less time and effort when interacting with chatbots (Mimoun et a., 2017).

Limitations of the present study and future work

There were five limitations to this study identified. One limitation of this study can be ascribed to the circumstances due to the ongoing COVID-19 pandemic, which did not allow for the experiment to be carried out in a lab setting. As a result, participants carried out the experiment at home on their own personal computers in an uncontrolled environment. Thus, there may have been unwanted distractors in their environment that disrupted the participants focus on the study. This problem was accommodated by supervising the participants through a Zoom conference, ensuring that participants would carry out the experiment in one session and enabling them to ask questions or for help when it was needed. Due to a relatively low turnout rate in participants, the option to carry out the experiment without supervision was given after three weeks of data collection. Although a manipulation test between supervised and unsupervised subjects did not suggest any difference in responses for both groups, there is still the possibility that there was a distractor in the environment of participants, supervised or not.

Furthermore, the sample population mostly comprised of German participants. Despite the option of a German translation of the survey, many German participants opted for the original English version. As many of the participants were students from the University of Twente, where speaking English is the standard for communication as well as education, it can be assumed that those participants who chose the English translation, despite English not being their first language, are used to this language and thus, also proficient in it.

Another limitation may have been the length of the task the participants had to carry out. While the shortened 15-item version of the Chatbot Satisfaction Scale puts less strain on participants compared to the 42-item version it is derived from, participants still had to interact with ten different chatbots which led to an expected length of 45-60 minutes to complete the experiment. Nevertheless, many participants completed the experiment in a shorter time of 25-35 minutes suggesting that they did not put enough effort or into the interaction with the chatbots and the subsequent filling out of the scales and possibly rushed while doing so.

It should also be considered that it proofed to be difficult to recruit participants who could be considered experts in the use of chatbots for this experiment. As full-time workers, the experts that were contacted were reluctant to participate in a supervised experiment that would last up to 60 minutes. The option of participating without supervision made it easier to recruit participants who are expected to have a higher level of familiarity with chatbots as they were better able to incorporate the participation into their schedule.

As it proved to be difficult to recruit experts, it may be possible that there was no correlation between familiarity and satisfaction found because none or only a few of the participants were familiar with chatbots to a level where they could be considered experts. Thus, conducting more research on the relationship between familiarity and high satisfaction scores may be reasonable if it is possible to recruit participants who are more familiar with chatbots. It may be helpful to experiment entirely unsupervised to allow for participants to be able to find time for their participation. Furthermore, carrying out the experiment with fewer chatbots may also be beneficial in this regard, as participants would need less time to complete the experiment and further making the participation more flexible. Asking participants to interact with fewer chatbots may also help in keeping participants patient so that they do not rush through the tasks. This may

also offer the possibility to give multiple tasks that participants have to perform with the chatbot, allowing the participants to interact with the chatbot for a longer time. Thereby, the items intended to assess the abilities of the chatbots to integrate contextual information may offer different results compared to the present study. The experiment may also appear as less of a burden to many due to it being shorter and attract more participants, correcting for the losses in data size when for example only 5 chatbots are interacted with.

Conclusion

This study showed that the Chatbot Satisfaction Scale works as a tool for the assessment of the perceived usability of chatbots. It showed that the already good model that was established through exploratory factor analysis can be improved by using fewer items without decreasing the reliability of the scale. A high correlation of the test scores from the Chatbot Satisfaction Scale was found when comparing it to the UMUX-Lite, supporting the concurrent validity of the scale. A German translation of the scale was also established which proved to be reliable and showed a moderately high correlation with the original English version of the scale. The option to offer the scale in more languages such as German will allow for broader use in the future. To conclude, this study showed that while already being a good assessment tool, further improvements can be made, and more research can be conducted in possible other aspects such as familiarity which may influence satisfaction ratings when measured in an adequate population.

References

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

Alba, J. W. (1983). The effects of product knowledge on the comprehension, retention, and evaluation of product information. *ACR North American Advances*.

Angga, P. A., Fachri, W. E., Elevanita, A., Suryadi, & Agushinta, R. D. (2015). Design of chatbot with 3D avatar, voice interface, and facial expression. *2015 International Conference on Science in Information Technology (ICSITech)*. doi:10.1109/icsitech.2015.7407826

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 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*. doi: 10.1016/j.cosrev.2020.100239

Berkman, M. I., & Karahoca, D. (2016). Re-assessing the usability metric for user experience (UMUX) scale. *Journal of Usability Studies, 11*(3), 89-109.

Bevan, N. (1995). Usability is quality of use. In *Advances in Human Factors/Ergonomics* (Vol. 20, pp. 349-354). Elsevier.

Borsci, S., Buckle, P., & Walne, S. (2020). Is the LITE version of the usability metric for user experience (UMUX-Lite) a reliable tool to support rapid assessment of new healthcare technology? *Applied Ergonomics, 84*. doi: 10.1016/j.apergo.2019.103007

Borsci, S., Federici, S., Gnaldi, M, Bacci, S., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: An exploratory analysis of SUS, UMUX, and UMUX-LITE. *International Journal of Human-Computer Interaction, 31*, 484-495.

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (Under review). The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents.

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

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

Brucks, M. (1985). The effects of product class knowledge on information search behavior. *Journal of consumer research*, *12*(1), 1-16.

Capgemini. (2019). *Smart talk: How organisations and consumers are embracing voice and chat assistants*. https://capgemini.com/wp-content/uploads/2019/09/Report_Conversational-Interfaces-1.pdf

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

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737-758. doi:10.1016/s1071-5819(03)00041-7

Davis, D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quaterly, 13*, 319-339.

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

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega*, *28*(6), 725-737.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behaviour, 97*, 304-316.

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *ICIS 2017: Transforming Society with Digital Innovation*.

Gnewuch, U., Morana, S., & Maedche, A. (2017, December). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *ICIS*.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction.

Huang, H., Cerekovic, A., Tarasenko, K., Levacic, V., Zoric, G., Pandzic, I. S., Nakano, Y., & Nishida, T. (2008). Integrating Embodied Conversational Agent Components with a Generic Framework. *International Journal of Multiagent and Grid Systems, 4*(4), 371-386.

International Organization for Standardization (2018). *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11). Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

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

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Lewis, J.R. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction, 34*(12), 1148-1156. doi:10.1080/10447318.2017.1418805

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., 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: Association for Computing Machinery.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction, 31*(8), 496-505.

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

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management review*, *23*(3), 473-490.

McTear, M., Callejas, Z., & Griol, D. (2016). Conversational interfaces: devices, wearables, virtual agents, and robots. In *The Conversational Interface* (pp. 283-308). Springer, Cham.

Mimoun, M. S. B., 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.

Mimoun, M. S. B., 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.

Miyake, N., & Norman, D. A. (1979). To ask a question, one must know enough to know what is not known. *Journal of verbal learning and verbal behavior*, *18*(3), 357-364.

Pamungkas, E. W. (2018). Emotionally-Aware Chatbots: A Survey. *Proceedings of ACM Conference.* doi:10.1145/1122445.1122456

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

Qualtrics. (n.d.) Retrived from http://www.qualtrics.com/

Qui, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management 8Information Systems, 25*(4), 145-182. doi: 10.2753/MIS0742-1222250405

R Core Team. (2021). R: A language and environment for statistical computing (Version 4.0.5). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Rafaeli, S. (1988). Interactivity: From new media to communication. In R. P. Hawkins, J. M. Wiemann, & S. Pingree (Eds.). *Advancing communication science: Merging mass and interpersonal processes* (pp. 110-134). Newbury Park, CA: Sage.

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

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3,.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Seeger, A. M., Pfeiffer, J., & Heinzl, A. (2017). When do we need a human? Anthropomorphic design and trustworthiness of conversational agents. *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS, AISeL, Soeul, Korea* (Vol. 10)

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

Silderhuis, I., & Borsci, S. (2020). *Validity and Reliability of the User Satisfaction with Information Chatbots Scale (USIC).* (Master thesis). University of Twente, Enschede, Netherlands.

Somasundaram, S., Kant, A., Rawat, M., & Maheschwari, P. (2019). *The future of chatbots in insurance*. Cognizant. https://www.cognizant.com/whitepapers/the-future-of-chatbots-in-insurance-codex4122.pdf

Tariverdiyeva, G. (2019). *Chatbot's Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis*. (Masters thesis). University of Twente, Enschede, Netherlands

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

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wang, W., & Benbasat, I. (2008). Attribution 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

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. https://CRAN.R-project.org/package=dplyr

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. doi: 10.1145/3125739.3125766

Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*(1), 53-93.

Zoom. (n.d.) Retrieved from https://zoom.us

Appendices

Appendix A: Five factor structure of the 15-item Chatbot Satisfaction Scale

Factor	Item
1. Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable.
	2. It was easy to find the chatbot.
2. Perceived quality of chatbot functions	3. Communicating with the chatbot was clear.
	4. I was immediately made aware of what information the chatbot can give me.
	5. The interaction with the chatbot felt like an ongoing conversation.
	6. The chatbot was able to keep track of context.
	7. The chatbot was able to make references to the website or service when appropriate.
	8. The chatbot could handle situations in which the line of conversation was not clear.
	9. The chatbots responses were easy to understand.
3. Perceived quality of conversation and information provided	10. I find that the chatbot understands what I want and helps me achieve my goal.
	11. The chatbot gives me the appropriate amount of information.
	12. The chatbot only gives me the information I need.
	13. I feel like the chatbot's responses were accurate.
4. Perceived privacy and security	14. I believe the chatbot informs me of any possible privacy issues.
5. Time response	15. My waiting time for a response from the chatbot was short.

Appendix B: Chatbot Satisfaction Scale (English original)

Chatbot Satisfaction Scale from Borsci et al. (under review)

Could be answered on a five-point Likert Scale ranging from: Strongly disagree (1) to Strongly agree (5)

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

Item	Description

1	The chatbot function was easily detectable.
2	It was easy to find the chatbot.
3	Communicating with the chatbot was clear.
4	I was immediately made aware of what information the chatbot can give me.
5	The interaction with the chatbot felt like and ongoing conversation.
6	The chatbot was able to keep track of context.
7	The chatbot was able to make references to the website or service when appropriate.
8	The chatbot could handle situations in which the line of conversation was not clear.
9	The chatbots responses were easy to understand.
10	I find that the chatbot understands what I want and helps me achieve my goal.
11	The chatbot gives me the appropriate amount of information.
12	The chatbot only gives me the information I need.
13	I feel like the chatbot's responses were accurate.
14	I believe the chatbot informs me of any possible privacy issues.
15	My waiting time for a response from the chatbot was short.

Appendix C: Chatbot Satisfaction Scale (German translation)

Die folgenden Fragen werden auf einer fünf-Punkt Likert Skala beantwortet von: Stimme überhaupt nicht zu (1) - Stimme voll und ganz zu (5)

Beantworten Sie die nächsten Aussagen anhand Ihrer Erfahrung mit dem Chatbot:

Item	Description
1	Die Chatbot-Funktion war leicht zu erkennen.
2	Es war einfach den Chatbot zu finden.
3	Die Kommunikation mit dem Chatbot war eindeutig.
4	Ich wurde sofort darauf aufmerksam gemacht, welche Informationen mir der Chatbot geben kann.
5	Die Interaktion mit dem Chatbot fühlte sich wie eine laufende Unterhaltung an.
6	Der Chatbot war in der Lage, den Kontext zu verfolgen.
7	Der Chatbot war in der Lage, bei Bedarf Verweise auf die Website oder den Service zu machen.
8	Der Chatbot konnte mit Situationen umgehen, in denen die Gesprächsrichtung nicht klar war.

9	Die Antworten des Chatbots waren einfach zu verstehen.
10	Ich finde, dass der Chatbot versteht, was ich will und mir hilft, mein Ziel zu erreichen.
11	Der Chatbot gibt mir die angemessene Menge an Informationen.
12	Der Chatbot gibt mir nur die Informationen, die ich brauche.
13	Ich habe das Gefühl, dass die Antworten des Chatbots korrekt waren.
14	Ich vertraue darauf, dass der Chatbot mich über mögliche Datenschutzprobleme informiert.
15	Meine Wartezeit auf eine Antwort des Chatbots war kurz.

Appendix D: UMUX-Lite (German translation)

Die folgenden Fragen werden auf einer fünf-Punkt Likert Skala beantwortet von: Stimme überhaupt nicht zu (1) - Stimme voll und ganz zu (5)

Beantworten Sie die nächsten Aussagen anhand Ihrer Erfahrung mit dem Chatbot:

Item	Description
1	Die Fähigkeiten dieses Systems erfüllen meine Anforderungen.
2	Dieses System ist einfach zu bedienen.

Appendix E: Chatbots and tasks

1. https://www.chatbot.com

Perform the following task using the chatbot:

You are interested in implementing a chatbot onto your website. You want to find out the price for the least expensive plan.

German description:

Sie sind daran interessiert, einen Chatbot auf Ihrer Website zu implementieren. Sie möchten den Preis für das günstigste Angebot herausfinden.

2. https://www.utwente.nl/en/education/master/chat/?autostart=true

Perform the following task using the chatbot:

You are a Dutch student who would like to do a Master's degree at the University of Twente. Your name is Jack/Jacky and when you are asked for your email you can decline this. You are interested in doing your master in Interaction Technology 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.

German description:

Sie sind ein niederländischer Student, der an der Universität Twente ein Masterstudium absolvieren möchte. Ihr Name ist Jack/Jacky und wenn Sie nach Ihrer E-Mail gefragt werden, können Sie dies ablehnen. Sie sind daran interessiert, Ihren Master in Nanotechnologie im September 2021 zu machen. Sie haben Ihren Bachelor an der UTwente in den Niederlanden gemacht. Sie fragen den UTwente-Chatbot, welche Möglichkeiten es für ein Stipendium gibt.

3. https://www.amtrak.com/home.html

Perform the following task using the 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.

German description:

Sie möchten von Boston nach Washington D.C. reisen während Sie in den USA sind. Sie möchten den Chatbot von Amtrak nutzen, um die kürzestmögliche Fahrt für den 8. Oktober zu buchen. Ihr Abfahrtsbahnhof is Back Bay Station.

4. https://www.lufthansa.com/digitalassistant/webchat.html

Perform the following task using the chatbot:

You want to re-book your flight which you bought after May 15 2020. You bought it directly with Lufthansa.

German description:

Sie möchten Ihren Flug, den Sie nach dem 15. Mai 2020 gekauft haben, umbuchen. Sie haben ihn direkt bei Lufthansa gekauft.

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

Perform the following task using the chatbot:

You visit the Emirates Holidays page and use Emirates Holidays' chatbot to book a honeymoon holiday from the 4th September until the 9th October to London for two persons.

German description:

Sie besuchen die Emirates Holidays website und benutzen den Chatbot um Flitterwochen vom 4. September bis zum 9. Oktober in London zu buchen.

6. https://www.hdfcbank.com/personal/ways-to-bank

Perform the following task using the chatbot:

You are new to online banking and would like to know what a SIP is.

German description:

Sie sind neu im Online-Banking und würden gerne wissen, was eine SIP ist.

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

Perform the following task using the chatbot:

You are interested in requesting a demo of their solutions for your website. You would like to know what form you need to fill in.

German description:

Sie sind daran interessiert, eine Demo der Softwarelösungen für Ihre Website anzufordern. Sie würden gerne wissen, welches Formular Sie ausfüllen müssen.

8. https://www.benefitcosmetics.com/en-us

Perform the following task using chatbot:

You are interested in buying a brown mascara. Find out what options there are.

German description:

Sie interessieren sich für den Kauf einer braunen Wimperntusche. Finden Sie heraus, welche Möglichkeiten es gibt.

9. https://www.voegol.com.br/en

Perform the following task using the chatbot:

You want to know which destination GOL fly to, you are interested in national destinations in the southern area.

German description:

Sie möchten wissen, welche Ziele GOL anfliegt. Sie interessieren sich für inländische Ziele im südlichen Bereich.

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

Perform the following task using the chatbot:

You are interested in finding out where the Absolut is from.

German description:

Sie möchten wissen, wo Absolut herkommt.

Appendix F: Informed consent form

English version

Consent Form

Taking part in the study
I have read and understood the study information. 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 me interacting with different chatbots. The whole experiment will take about 60 minutes. I understand that for participating in the study there are no known risks involved**.** I am at least 18 years old.

Use of the information in the study
I understand that taking part in the study involves answering questions about my demographics, performing tasks and interacting with chatbots online and filling out two scales about each of the chatbots I have interacted with online.

Future use and reuse of the information by others
I understand that information I provide will be used for a bachelor thesis. I understand that before the information is achieved it will be anonymized by removing name and other information that could track me back. 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 research and learning.

Contact Information for Questions about Your Rights as a Research Participant
If you ever have any questions after this session has ended you can email us: s.m.kerwienlopez@student.utwente.nl or n.pollmann-1@student.utwente.nl and our supervisor can be reached at s.borsci@utwente.nl. 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

German version

Einverständiserklärung

Teilnahme an dieser Studie
Ich habe die Informationen zu dieser Studie gelesen und ich habe sie verstanden. Ich willige freiwillig ein an dieser Studie teilzunehmen und verstehe, dass ich ohne einen Grund zu nennen jederzeit die Möglichkeit habe auszusteigen. Mir ist bewusst, dass ich durch die Teilnahme an dieser Studie mit verschiedenen Chatbots interagieren werde. Das Experiment wird ungefähr 60 Minuten dauern. Mir ist bekannt, dass mit der Teilnahme an dieser Studie keine bekannten Risiken verbunden sind**.** Ich bin mindestens 18 alt.

Verwendung von Informationen in dieser Studie
Mir ist bewusst, dass die Teilnahme an dieser Studie die Beantwortung von Fragen zu meinen demographischen Daten, die Durchführung von Aufgaben und die Interaktion mit Chatbots online sowie das Ausfüllen von zwei Skalen über jeden der Chatbots beinhaltet.

Künftige Nutzung und Wiederverwendung der Informationen durch andere
Mir ist bekannt, dass die von mir zur Verfügung gestellten Informationen für eine Bachelorarbeit verwendet werden. Ich verstehe, dass die Informationen vor der Verwendung anonymisiert werden, indem Name andere Informationen, die mich zurückverfolgen können, entfernt werden. Ich erteile die Erlaubnis, dass das Ausfüllen der Skalen und des demographischen Fragebogens, den ich zur Verfügung stelle, in einem sicheren Datenspeicher archiviert, damit es für zukünftige Forschung und Lernzwecke verwendet werden kann.

Kontaktinformationen für Fragen zu Ihren Rechten als Studienteilnehmer
Wenn Sie nach dieser Sitzung noch Fragen haben, können Sie uns eine Email schicken: s.m.kerwienlopez@student.utwente.nl oder n.pollmann-1@student.utwente.nl
Mein Supervisor ist mit der folgenden Email-Adresse zu erreichen: s.borsci@utwente.nl
Wenn Sie Fragen zu Ihren Rechten als Studienteilnehmer haben oder Informationen einholen, Fragen stellen oder Bedenken zu dieser Studie mit einer anderen Person als dem/den Forscher(n) besprechen möchten, wenden Sie sich bitte an das Sekretariat der Ethikkommission der Fakultät für Verhaltens-, Management- und Sozialwissenschaften der Universität Twente unter:
ethicscommittee-bms@utwente.nl

Appendix G: Questions about familiarity

Familiarity

Could be answered on a five-point Likert scale ranging from: Not familiar at all (1) - Extremely familiar (5)

Item	Description
1	How familiar are you chatbots and/or other conversational agents?
2	How familiar are you with the way chatbots and/or other conversational agents work?

Chatbot Use

Could be answered on a five-point Likert scale ranging from: No (1) - Yes (5)

Item	Description
3	Have you used a chatbot or a conversational interface before?

Could be answered on a six-point Likert scale ranging from: Never (1) - Daily (6)

Item	Description
4	How often do you use chatbots and/or other conversational interfaces?

Confidence

Could be answered on a six-point Likert scale ranging from: Not confident at all (1) - Extremely confident (5)

Item	Description
5	How confident do you feel using a chatbot and/or conversational interface?