03/07/2021

User satisfaction and trust in chatbots: testing the Chatbot Usability Scale and the relationship of trust and satisfaction in the interaction with chatbots

Bachelor Thesis

Alina Waldmann
S2132893

First supervisor: dr. Borsci
Second supervisor: prof. dr. van der Velde

University of Twente
BMS Faculty
Department of Psychology

Abstract

Chatbots are rising in popularity and are commonly implemented in the customer service domain in recent years. Research about the motivations for users to interact with the chatbots was frequently conducted, which resulted in many aspects that show importance in user experience with chatbots. Two components of user experience are satisfaction and trust, however, there was no universal questionnaire to test for any of the two. The Chatbot Usability Scale (CUS) is a new scale which was specifically developed for user satisfaction evaluations with chatbots. Moreover, trust can have an influence on the user experience with technologies, either initially or by influencing the usage continuation intention. The present study aimed to test the psychometric properties of the a new scale, and investigated the relationship between initial trust, satisfaction, trust after the interaction, and usage continuance intention. Furthermore, the Usability Metric for User Experience Lite (UMUX-LITE) was used to check for external validity of the CUS.

A study with forty participants was conducted in which each participant tested and evaluated ten different chatbots based on the CUS, the UMUX-LITE, and questions regarding initial trust, trust after the interaction, and usage continuance intention.

A confirmatory factor analysis was performed to test for the psychometric properties of the CUS for the current sample. Results could not confirm the initial five-factor structure of the CUS, as the given sample showed poor model fit. Furthermore, correlation analyses between the CUS and the UMUX-LITE were conducted, which showed good correlations with $r = 0.804$ ($p< 0.001$) for the initial CUS and $r = 0.809$ ($p< 0.001$) for the modified model. The relationship between initial trust, trust after the interaction, satisfaction and usage continuation intention was tested with a linear mixed-effects model. Results suggested that there was an effect of ‘Personal Innovativeness’ on ‘Satisfaction’ rated by the new scale, and that ‘Trust after the Interaction’ and ‘Satisfaction’ are affecting each other. Lastly, both ‘Satisfaction’ and ‘Trust after the Interaction’ seem to affect the ‘Usage Continuance Intention.’

Keywords: Chatbots, user experience, user satisfaction, usability, trust, initial trust, Chatbot Usability Scale, CUS, UMUX-LITE

Acknowledgements

First, I would like to thank my supervisor dr. Simone Borsci for his patience, kindness, and support he was showing me throughout the semester.

Furthermore, I would like to thank my family and friends but especially my mom for always being there for me whenever I needed support or did not know how to go on. Lastly, many thanks to my friend who took his time for helping me in the final refinements of the report.

Contents

List of tables

1. Introduction

Chatbots are conversational user interfaces, which understand and use natural language for interacting with the user in a text-based way. They are rising in popularity in recent years (Dale, 2016) although already in the 1960s, Joseph Weizenbaum's ELIZA was developed as one of the earliest chatbots. ELIZA was based on a set of stored patterns that matched against the user's input (Gwenuch et al., 2017). Thus, its options to interpret and react to the users' input were limited to what Weizenbaum has programmed into the software. Recent artificial intelligence (AI) advancements, by contrast, make chatbots capable of natural language processing and machine learning (Gnewuch et al., 2017; Skjuve & Brandtzæg, 2019). Hence, conversations can be more complex and more varied today than several years ago. Another reason why chatbots' popularity has increased might lie in the way people communicate with each other compared to people in recent years. As people are using email or chatrooms on the internet (Jenkins et al., 2007), or as 7.3 billion people use an SMS-capable mobile phone in 2017 (Dale, 2017), interacting with each other in a text-based way is more common. Based on these developments, chatbots became useful in various directions such as customer service agents, health advisors, therapists, or teachers (Skjuve & Brandtzæg, 2019).

Especially the usefulness of chatbots in the domain of customer services is of high interest for organizations (Gnewuch et al., 2017). Here, conversational interfaces are claimed as time-saving, fast, convenient, and cost-effective, but still give the company the chance of playing an active part during the interaction with the customers (Gnewuch et al., 2017; Jenkins et al., 2007). However, most of the chatbots so far could not meet the expectations of the customers and disappeared (Gnewuch et al., 2017). Several studies have evaluated the reasons why customers would not use a chatbot in customer service. As one of the greatest factors, users are unsatisfied with the skills the chatbots have (Kvale et al., 2021). Hereby, chatbots easily get out of context (Nuruzzaman & Hussain, 2018), give nonsensical answers (Brandtzæg & Folstad, 2017), or have problems understanding the users' questions (van der Goot et al., 2021). Reasons for this are that chatbots do not recognize grammatical errors (Nuruzzaman & Hussain, 2018), which let them have trouble, for example, to correctly interpret what the user wants to say if there are misspellings in the question. Moreover, current chatbots are not able to detect emotions that users have during the conversation (Nuruzzaman & Hussain, 2018), although the adequate responding to the mood or tone of voice of the user has found to be essential for customer service chatbots, as it influences the whole conversation experience (Kvale et al., 2021; van der Goot et al., 2021). Consequently, reasons for customers not to use chatbots especially lie in their limited skills to adequately directing the users to their goal (Kvale et al., 2021). This shows that, despite the recent developments in AI, chatbots are still facing problems in having an efficient and ongoing conversation using natural language.

Looking at all these rather fundamental limitations chatbots display, one may ask the question why the usage of chatbots is justifiable and useful at all. Følstad and Skjuve (2019), who investigated motivational factors of why users are using chatbots in customer services, gave an answer to this. They found that users are very much aware of the fact that chatbots only have limited capabilities, mostly including questions that can be answered in a straightforward manner. In order to overcome these communication issues, users adapt their behavior to this by formulating simple questions or sentences rather than telling the chatbot the whole problem in a detailed text (Følstad & Skjuve, 2019). Additionally, other users are using keywords right away in order to keep the conversation as simple as possible (van der Goot et al., 2021). Nevertheless, users are motivated to use chatbots because they can provide efficient and fast support (van der Goot et al., 2021). Hereby, users can get simple and easy to understand information without having to read through lots of text or pages of the website until finding what they are actually searching for (Følstad & Skjuve, 2019). Therefore, it can be time saving to ask a chatbot right away. Lastly, the aspect of availability is found as being a motivational factor for using chatbots in customer service. Chatbots are available whenever the customer is needing them. Conclusively, users do not have to wait in line or for the customer service to open, but can get information all the time (Følstad & Skjuve, 2019). This shows that customers seem to indeed value the option of having a chatbot, even if it still is beneficial to further develop chatbots in terms of ability to also solve more complicated questions in the long run (Følstad & Skjuve, 2019).

Other than being able to answer more complicated questions, Gnewuch et al. (2017) claim that the way of communication with the customer is also of great importance. As Nass et al. (1994) found in their research, people are unconsciously associating human characteristics to technological interfaces. Hence, users apply social norms, such as politeness, to the technological interface, expect it to adhere to them, and evaluate the interface based on them afterwards (Nass et al., 1994). Thus, technological interfaces are expected to communicate in a way that is in accordance with human characteristics. Thereby, they are treated like social actors, which, subsequently, should pertain to the interaction with chatbots as well.

Jenkins et al. (2007) were also focusing on the conversation users have with a chatbot and what they are expecting from it, respectively. Additionally to Nass et al.'s (1994) expectation of human characteristics, Jenkins et al. (2007) found that productivity is also of great importance for a chatbot. More concrete, users expect that the chatbot should be able to provide information in a shorter amount of time than a human (Jenkins et al., 2007). Hence, Jenkins et al. (2007) conclude that the chatbot system should establish a rapport with the customer by having the same tone, sensitivity and behavior as a human (human characteristics), while giving the user the information they are interested in and guide them to the right parts of the website (productivity).

However, the term *human characteristics* can be defined very broadly in the sense of (chatbot) interactions. Gnwuch et al. (2017), therefore, were searching for specific factors that might be important for the impression of the conversation. Next to merely chatbot related aspects such as quantity, quality, relation and manner, they also found social factors to be of influence on the perceived quality of a conversation. Gnwuch et al. (2017) refer to the Social Response Theory's physical, psychological, and language factors as well as its factors of social dynamics and social roles, and conclude that those are of value for the perceived quality of a conversation with a chatbot. However, a further specification of what these latter characteristics imply in the context of human-computer interaction (HCI) was not given.

Another reason for having a closer look into humanlike cues of chatbots is that human likeness can positively influence relationship-building, not only between the chatbot and the user but especially between the customer and the company (Araujo, 2018). As companies are dependent on satisfied customers, it should, therefore, be important for them to reflect the different channels that influence the company's image, such as through the chatbots they provide on their websites. However, participants of van der Goot et al.'s (2021) research expressed that they often have the feeling that chatbots are rather implemented for the company itself (i.e. to save resources for instance), than for the advantage of the customers. Accordingly, the results of Araujo (2018) may represent important implications for the companies in regard to their chatbots: users feel more emotionally connected to the company when the chatbot presented humanlike cues. Hereby, humanlike language or a name for the chatbots are already enough for this to happen (Araujo, 2018). Conclusively, a lot of characteristics influence the experience of the user regarding the chatbot but the companies should take these aspects seriously as their decisions regarding the chatbot may also have an effect on the company itself.

1.1 User experience

As all this shows, many aspects influence the perception and motivations users have regarding chatbots and its use, which, consequently, have an impact on the users' impression of the company. Hence, *user experience* is a key aspect of successfully implementing chatbots in customer services as an advantage for the company. User experience, as defined by the international standard of human-centered design, ISO 9241-210, are the "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" (cited in Følstad & Brandtzæg, 2020, p.3). In a recent study by Kvale et al. (2021), they investigated different aspects that are influencing user experience with chatbots. As a means, they used customer satisfaction surveys to analyze differences regarding satisfaction affected by problem-solving attainment, the kind of problem, and characterized intents associated with positive and negative user experiences. However, Kvale et al. (2021) mention that, although the customer satisfaction survey may provide a valuable reflection of good or poor user experience, it is not sufficient for providing the needed nuances for an overall user experience construct. Nevertheless, satisfaction is found to be a valuable aspect when evaluating interfaces, as satisfaction is furtherly evaluated on the aspects of usability and usefulness of a system (Tsakonas and Papatheodorou, 2008). Usefulness can be defined as the value a tool has for the completion of a task, while usability is about the "effective, efficient and satisfactory task accomplishment" (Tsakonas & Papatheodorou, 2008, p.1238). Altogether, these aspects are all part of the user experience (Følstad & Brandtzæg, 2020), thus, satisfaction might still be a good means for evaluating it.

Other means to explore the dimensions of usability and satisfaction than customer satisfaction surveys are given by short scales, i.e. short satisfaction questionnaires. One of the most popular scales used for this background is the System Usability Scale (SUS; Borsci et al., 2015; Lewis, 2006). It is a ten-item scale that showed high reliability and demonstrated validity (Borsci et al., 2015). Two other even shorter, i.e. ultrashort, scales are the Usability Metric for User Experience (UMUX; Finstad, 2010), and the UMUX-LITE (Lewis et al., 2013). While the former consists of four items, the latter has only two items, while both still show to be reliable (Borsci et al., 2015). Consequently, all three scales are likely used for satisfaction evaluations of websites as they need little time by still providing valuable results. In their research, Borsci et al. (2015) compared the SUS, UMUX and UMUX-LITE with each other and showed that all scales show a high correlation with each other. This means that all three are measuring the same underlying construct of satisfaction and are equally valuable for satisfaction evaluation. Thus, these scales could provide a good means for companies to effectively and efficiently evaluate their chatbots by customers. However, van den Bos and Borsci (2021) point out that these usability scales are not made for evaluating interactive interfaces, such as chatbots. Chatbots hold a set of particular characteristics that are more diverse than other user-technology interactions (Følstad et al., 2018), which are not included in these scales to sufficiently measure user-chatbot interaction (van den Bos & Borsci, 2021). Hence, although these scales are valuable for evaluating satisfaction on websites, for example, they do not bring sufficient results in the context of chatbots in customer services.

However, there are more features than usefulness, usability, and satisfaction related to user experience in user-chatbot interaction. In order to get to know these aspects, Følstad and Brandtzæg (2020) applied a questionnaire study in which participants evaluated their chatbot experiences. In the end, their findings reflect above already mentioned factors, like *help and assistance*, or *social and human likeness* (Følstad & Brandtzæg, 2020). So, there have already been given a lot of characteristics that chatbots should present in order to improve the user experience during the interaction and the eventual evaluation of and satisfaction with it. Nevertheless, a scale that implies the most important of them into one scale measuring satisfaction in chatbots was only recently developed (Borsci et al., Under Review; see section 1.3).

1.2 Trust

Next to satisfaction, usability, and usefulness as dimensions of user experience, *trust* is also found to be a factor of influence on the users' experience with the system (Følstad & Brandtzæg, 2020). Trust is a subject of ongoing research for decades already but with no universally accepted definition so far. A difficulty in this is that many different factors can influence trust, depending on the specific context the person is in while trusting (McKnight & Chervany, 1996). Commonly, trust is defined as the "individual's willingness to depend on another party because of the characteristics of the other party" (McKnight et al. 2011, p.12:1). Furthermore, Rosseau et al. (1998) define trust as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (p.395). Comparing these two definitions, there seem to be two central aspects that influence trust: the dependence on the other person, and the characteristics of the other person one is interacting with. Wang and Emurian (2005) name this 'other person' the *trustee*, hence, the party which one is dependent on while trusting, while the one who trusts is the *trustor*. In the context of this study, the trustor is the user/customer, and the trustee is the chatbot.

Trust is seen as essential in human relationships and, thereby, represents a central aspect of how we interact with others (McKnight & Chervany, 1996; Wang & Emurian, 2005). Therefore, it is studied in several domains such as philosophy, psychology, and marketing (Wang & Emurian, 2005). Moreover, trust is found to also be important in the context of technology as it influences individual decisions to use that specific technology (Følstad et al., 2018; McKnight et al., 2009). An area in which trust has widely been researched in the context of technology is in the area of eCommerce (e.g. Gefen & Straub, 2004; McKnight et al., 2002), which is about the buying and selling of goods or services on the internet. Here, it is found that trust has a strong effect on purchase behavior (Gefen & Straub, 2004), which, therefore, represents the importance of trust in a company-customer relationship. As chatbots in customer services are also related to eCommerce in that regard that their help can influence the purchase behavior of the customer, they can also be seen as an important means for a trusting relationship with the company. However, due to their highly particular characteristics, research on eCommerce cannot be reflected in an one-on-one relationship to chatbots (Følstad et al., 2018). However, putting together the rising popularity of chatbots, the importance of trust in a relationship and its effect on the behavior of the trustor regarding use and outcome, it should be suggested that trust also plays a crucial role in the implementation of chatbots. Nevertheless, research in this area is rather rare although there seems to be a rising interest by now (Følstad et al., 2018).

1.2.1 Initial trust and usage continuance intention

Trust can have an influence on the user experience with the technology or chatbot respectively, in several stages of the interaction. At the beginning, there is *initial trust* (McKnight et al., 1998). Initial trust is characterized by the fact that the trustor is interacting with an unfamiliar trustee, meaning that both of them have not made a meaningful bond with each other yet (McKnight et al., 2002). Consequently, it implies the amount of trust a trustor gives a new and yet unknown trustee. This stage of trusting is arguably seen to be the most influential in the relationship-building between two parties (McKnight et al., 1998). This as a ground, McKnight et al. (1998) have created a model of how trust is built by the influence of several aspects. Hereby, they distinguish between a disposition to trust, i.e. the consistent tendency to trust throughout different situations and persons (McKnight & Chervany, 1996), institution-based trust, i.e. the individual's perception of the institutional environment

(McKnight et al., 2002), and cognitive processes, that eventually influence the development of trust (McKnight et al., 1998). To sum it up, the initial tendency of a person to trust another person, the environment in which they are in and related cognitive processes were found to influence the impression of a trustor's trust level towards a trustee.

In a later study about initial trust and its relation to consumer adoption of eCommerce, McKnight et al. (2002) enhance the model of McKnight et al. (1998) and test if the theoretical framework can also be confirmed in statistical research. In the end, they could validate their model, showing that four higher level interrelated trust constructs can model the development of trust in eCommerce: disposition to trust, institution-based trust, trusting beliefs, i.e. the perceptions of the trustee's attributes, and trusting intention, i.e. the intention to engage in trust-related behaviors (McKnight et al., 2002). Moreover, disposition to trust was found to have a significant effect on personal innovativeness, i.e. how much the person is interested in exploring new technology, but it was not furtherly investigated whether personal innovativeness also has an effect on the eventual trust-related behavior. Consequently, many aspects of initial trust were found to impact trusting relationships with eCommerce, which may also play a role in user-chatbot interactions as tested in the current study.

Next to initial trust, i.e. trust before the trustor knows the trustee, the influence of trust is also operating over time. Lankton et al. (2014) investigated trust in technology regarding its influence on two aspects: satisfaction, i.e. how pleased the user was with the device, and usage continuance intention, i.e. if the user would be willing to use the device further. Especially the latter should be of great interest for companies as they want their product to be bought and used. For this, they made a two-part study by first investigating the initial trust towards a system, and subsequently testing for the development of trust toward this system six weeks later. They found that trusting intention significantly influences usage continuation and predicts it better than satisfaction does. However, trust does not affect satisfaction but satisfaction affects trust. Conclusively, satisfaction influences trust which in turn influences usage continuation. This is an important implication for companies as they want their products to be bought and used. In the context of chatbots in customer services, this finding might be meaningful as companies use chatbots to limit their resources, but Lankton et al. (2014) predict that users would only continue to use the chatbot if they are satisfied with it and, in turn, trust it. This furthermore shows the dimensionality of user experience as already mentioned above, as several aspects are influencing the outcome, which all should be taken into account when implementing a chatbot.

1.3 Chatbot Usability Scale (CUS)

As described in the last paragraph, user satisfaction with a system is of great importance for the implementation of that system, as it affects trust which, again, increases the usage continuance intention (Lankton et al., 2014). However, chatbots are much more diverse than other technologies, thereby showing a range of specific characteristics which cannot be measured by traditional questionnaires such as the SUS (Følstad et al., 2018; van den Bos & Borsci, 2021). That as a ground, Balaji and Borsci (2019) started to develop a questionnaire that should be applicable for satisfaction measurement in the context of chatbots. As a final version, they presented a 42-items questionnaire named *User Satisfaction with Information Chatbots* (USIC).

However, a 42-items satisfaction questionnaire is very time and energy consuming for the users to fill out. Hence, Borsci et al. (Under Review) conducted an exploratory factor analysis to detect factors and items to eventually shorten the USIC, which would put less strain on the user to fill out. Subsequently, they came up with a 15-items solution, testing five different underlying structures. An overview of the *Chatbot Usability Scale* (CUS) can be found in Table 1. Compared to the USIC, the CUS, therefore, has many advantages although it is very up to date. Therefore, this study will use the CUS for further testing.

Table 1

15-item Chatbot Usability Scale (CUS) developed by Borsci et al. (Under Review).

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 chatbot's 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.

1.4 Aim of this study

Borsci et al. (Under Review) recently developed the Chatbot Usability Scale (CUS), which aims to test the satisfaction of users with chatbot interactions. Thereby, they came up with a 15-items solution loading on five factors while having good reliability with a Cronbach’s alpha of 0.8. Since previous work conducted exploratory factor analyses, the present study aims to apply confirmatory factor analysis to test the psychometric properties of the 15-items questionnaires. Hence, the first research question is:

RQ1: Can the psychometric properties of the CUS be confirmed with a factor loading on five factors and a reliability over 0.7?

Further, due to the newness of the CUS, external validation has to be made so that the measurement scale can be generalized for broader populations as well. For this, the UMUX-LITE by Lewis et al. (2013) is a standardized two-items questionnaire for measuring user satisfaction with systems, that is applicable to investigate whether the CUS and the UMUX-LITE measure the same underlying concept. This, again, is an indicator for good external validation. Therefore, the second research question is:

RQ2: Does the 15-items CUS correlate with the UMUX-LITE?

As McKnight et al. (2002) point out, initial expectations towards a system are important in order to predict satisfaction. However, Lankton et al. (2014) also point out that initial expectations might also be influenced by the eventual experience with the system, which, furthermore, influences the feeling of trust as well. Hence, the initial trust, trust after the interaction, and satisfaction should be correlated with each other. Therefore, this study aims to test if initial trust (McKnight et al., 2002) affects satisfaction measured through the CUS (Borsci et al., Under Review) and Trust after the interaction with the chatbot (Lankton et al., 2014).

Due to time and cognition strain minimization for the participants, it was decided to only use parts of the model about initial trust by McKnight (2002). In the model by McKnight et al. (1998), *disposition to trust* was found to be the influencing factor for trusting beliefs and trusting intention which could partly be confirmed by McKnight et al. (2002). Furthermore, Gefen and Straub (2004) have found that disposition to trust has a significant effect on trust. Hence, it was decided to concentrate on this aspect of initial trust in this research. Moreover, McKnight et al. (2002) have found that disposition to trust has an effect on personal innovativeness but has not investigated whether personal innovativeness has an effect on trust as well. As chatbots are not that long applied by now, this study wants to investigate whether *personal innovativeness* has an effect on satisfaction and trust after the interaction (McKnight et al., 2002). Due to simplicity, the term *initial trust* is following used to cover these two aspects found by McKnight et al. (2002). Based on the same reasons as mentioned above, the questions used by Lankton et al. (2014) were only partly been used for this study. Consequently, *trust after the interaction* implies the aspects *technology trusting performance*, *technology trusting intention*, and *usefulness*. Therefore, the third research question is:

> RQ3: Do disposition to trust and personal innovativeness have an effect on satisfaction and trust after the interaction with the chatbot?

Moreover, trust and satisfaction are found to be related with each other as satisfaction also affects trust (Lankton et al., 2014). Thus, having high satisfaction should enrich the feeling of trust towards that chatbot. With the newly developed CUS, it is now possible to test for this relationship in user-chatbot interactions as well. However, it would also be of interest if trust after the interaction may affect satisfaction in a user-chatbot context. Hence, the fourth research question is:

> RQ4: Does satisfaction with chatbots have an effect on trust after the interaction and vice versa?

Lastly, Lankton et al. (2014) also emphasize that trust better predicts usage continuance intention than satisfaction. Thus, usage continuation should merely be determined by the trusting performance but less by the satisfaction itself. However, chatbots differentiate themselves from other technologies due to their specific characteristics for interaction with the user through natural language (Følstad et al., 2018). Therefore, it would be interesting to see if this uniqueness of chatbots affects the relationship between satisfaction and usage continuance intention, as it was identified by Lankton et al. (2014), or if trust still is the better predictor for usage continuation intention. Therefore, the fifth research question will be:

> RQ5: Is usage continuation intention affected by trust after the interaction and satisfaction measured by the CUS?

2. Methods

2.1 Participants

Through snowball sampling, 40 volunteers participated in the present study (M_{age}=29.40; SD_{age}=14.11). The age range was between 18 and 78 years old, and there was an equal number of female and male participants. The majority of the participants (87.5%) were German, while 10% were Dutch and one participant was a Macedonian citizen. Of the participants, 5% were extremely or very familiar with chatbots respectively, while 45% were moderately familiar, 35% slightly familiar and 10% of the participants were not familiar with chatbots at all. Moreover, almost half of the participants (47.5%) have definitely used a chatbot before, while 30% have probably used it, and 2.5% were unsure. 10% each have probably not or definitely not used a chatbot before respectively. Lastly, participants were asked about how frequently they use chatbots. Hereby, 77.5% of the participants indicated that they rarely use a chatbot, and 17.5% never do. One participant uses it four to six times a week, whilst another one uses chatbots two to three times a week. The overall experience of the participants with the companies used in this research was low (M_{exp}=1.705; SD_{exp}=1.160).

The research was approved by the Ethics Committee of the BMS faculty of the University of Twente. Before participating, participants read an information sheet and agreed with the informed consent (see Appendix A). Additionally, Psychology and Communication Science students from the University of Twente could earn course credits if they signed up through the corresponding system.

2.2 Materials

Qualtrics (n.d.) was used to gather data using an online questionnaire. Within Qualtrics (n.d.), four different aspects were investigated. First, initial trust was measured previous to the chatbot interactions by McKnight et al.'s (2002) disposition to trust and personal innovativeness questions (see Appendix B1). The former consisted of nine questions split up into three sub-categories with three questions each. First, there were questions about Benevolence, which asked about how much the participant think that people care about the well-being of other people. Secondly, three items of Integrity asked about how much the participants think that people are honest and keep their promises. The last three questions of disposition to trust were about Trusting Stance, hence asked about the participants' way of trusting other people. Furthermore, questions about personal innovativeness included five items that asked about the behavior of the participants regarding the exploration of new websites or technologies. In the following, both disposition to trust and personal innovativeness will be meant by the term *Initial trust*.

As a second part, user satisfaction was investigated after the interaction with the chatbots by the use of two scales: the 15-items CUS (see Appendix B2) developed by Borsci et al. (Under Review) and the two-items UMUX-LITE (see Appendix B3) developed by Lewis et al. (2013).

Third, questions about trust after the interaction with the chatbot were asked. Hereby, Lankton et al.'s (2014) questions about technology trusting performance, technology trusting intention, and usefulness were used (see Appendix B4). First, technology trusting performance asked nine questions regarding functionality, helpfulness, and reliability of the chatbot. Questions about technology trusting intention implied four items regarding how much the participants feel that they can rely on the capabilities of the chatbot. Lastly, four questions about the usefulness of the chatbot were asked. Due to simplicity, these three sub-scales will following be summarized by the term *Trust after the interaction*.

Lastly, three items were included that investigated Usage continuance intention by questions of Lankton et al. (2014; see Appendix B5). It was capturing the participants' willingness to continue using the chatbot in the future.

As the CUS was based on a 5-point Likert scale, it was decided to apply this to all other questionnaires as well, although the UMUX-LITE, Trust after the interaction and Usage continuance intention originally use 7-point Likert scales.

Moreover, Qualtrics (n.d.) was used to present the chatbots and tasks to the participants. Some of them were included from a previous study by van den Bos and Borsci (2021) while others were changed or newly included, however, all are used in the domain of customer service (see Appendix C). Due to the Covid-19 pandemic, Google Meets (n.d.) was used to have online meetings with the participants.

2.3 Task

The task was to firstly indicate whether the participant had experience with the presented company before. Beneath that question, participants found the scenario for which they subsequently would be using the chatbot (see Appendix C). Following the link to the website, they first had to find the chatbot and interact with it to, eventually, achieve the goal of the scenario. If the participant felt that they had completed the task or got the information they needed respectively, they went back to the questionnaire to proceed with the survey by filling out the scales of the CUS, UMUX-LITE, Trust after the interaction and Usage continuance intention. In the end, all participants had to interact with ten different chatbots following this same scheme.

2.4 Procedure

Participants received a link to Google Meets (n.d.) up to 24 hours before the start of the session together with the note to be as rested as possible, and in a quiet closed room before entering the meeting. When the participants entered the meeting, the researcher would welcome them, and ask the participant to switch their phone into flight mode and to put it away from them to prevent distractions by it. Furthermore, the researcher asked the participants whether they would be willing to share their screen to better be able to follow the participants' progress as well as being more efficient for answering questions. After arranging this, the researcher shared the Qualtrics (n.d.) link to the questionnaire and explained the main goal and task of the survey to the participant. After that, the researcher muted herself and turned off her camera, while participants could read the information sheet. Next, they actively had to agree on the consent form. If they ticked 'yes' on the question about recording the session, the researcher would start the recording. With eventually clicking on the arrow, the participants continued to the demographic questions, questions regarding their previous chatbot experience, as well as questions regarding Initial trust (McKnight et al., 2002). When the participants had filled out each, they arrived at the starting point of the chatbot testing. Before continuing, participants were informed that the following will not measure their ability to interact with a chatbot but their satisfaction with it alone. Furthermore, the procedure of the tasks was explained. If the participants felt that they understood everything, they could continue at their own speed and perform the tasks with each of the ten chatbots. This implied, that the participants had to find and use ten different chatbots on customer service websites, followed by answering the CUS, the UMUX-LITE, and questions about Trust after the interaction and Usage continuance intention (Lankton et al., 2014). For all scales used in the session the items were randomized each time. Meanwhile, the researcher stayed in the session, so that any questions or insecurities could be answered. When the participants were finished with the tasks, the researcher thanked the participant for participation, asked if there were any questions and how it went for the participant.

2.5 Data Analysis

Data was exported from Qualtrics (n.d.) to Microsoft Excel 365 in numeric values. In Excel, unnecessary columns of data were removed, and labels were given to the items. Subsequently, the data was rearranged so that there was one data line per chatbot and participant combination, which led to 400 data lines in total. Afterwards, the data was imported into R (v4.1.0; R Core Team, 2021) for further analysis.

2.5.1 Confirmatory factor analysis

As implied in the first research question, Borsci et al. (Under Review) have found factor loadings on five factors for the CUS (Table 1). This psychometric property should be confirmed during this research. It was decided to test this by performing confirmatory factor analyses. For this, the R package 'lavaan' by Rosseel et al. (2021) was used, whereby the parameters of M1 were specified according to the original CUS (Table 1).
First, the assumption of normality was assessed, using the Shapiro-Wilk test for each variable of the CUS. The evaluation criteria was set by $p_{norm}> 0.05$ to determine normal distribution (Hanusz et al., 2014). As these criteria were not fulfilled, model estimation was based upon the robust maximum likelihood (MLR; Li, 2016). Model fit was assessed by Chi-square goodness of fit statistics (χ^2), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and Comparative Fit Index (CFI). Conventional cutoffs were used to determine acceptable fit, including $p_{\chi2}> 0.05$; RMSEA$\leq$ 0.08 for acceptable fit (< 0.06 for good fit); SRMR< 0.05 for good fit; and CFI$\geq$ 0.95 (Barney et al., 2021; Harerimana & Mtshali, 2020). Furthermore, Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) were used as tools to select the best model. The model with the lowest AIC or BIC was indicated as the best model (Barney et al., 2021; Wang & Liu, 2006). Furthermore, reliability was measured using the 'Psych' package (Revelle, 2019). In the end, the R package 'semPlot' by Epskamp et al. (2019) was used for visualization of the final model.

2.5.2 Correlation analysis CUS and UMUX-LITE

To answer the second research question and explore the relationship between the CUS and the UMUX-LITE, a correlational analysis with Spearman's rank-order correlation was performed. For that, mean scores were computed for each row regarding the items included in the UMUX-LITE and CUS, which were individually compared with each other afterwards.

2.5.3 Regression analyses

For answering the third, fourth, and fifth research question, a linear mixed-effects model using the R package ‘nlme’ (Pinheiro et al., 2021) was performed with significance levels of p≤ 0.05. A linear mixed-effects model was used as it shows advantages regarding sample structure and independence of the measurements (Yang et al., 2014). A traditional linear regression model has the assumption of independence, which was hypothesized to not be confirmed by the current data; participants were measured repeatedly as everyone had to interact with ten chatbots, which is an indication of dependent measurements.

First, individual mean scores were examined for each row of data including the items of interest for ‘Disposition to Trust’, ‘Personal Innovativeness’, ‘Satisfaction’, ‘Trust after the Interaction’, and ‘Usage Continuance Intention’ respectively. Following, the data was standardized and fitted. Assumption testing was found to be acceptable for all variables, although not perfect for the variables regarding ‘Usage Continuance Intention.’ In the end, the R package ‘ggplot2’ (Wikham et al., 2021) was used for a visual representation of the significant relationships.

For the third research question regarding the influence of disposition to trust and personal innovativeness on satisfaction and trust after the interaction, ‘Disposition to Trust’ and ‘Personal Innovativeness’ were treated as independent variables affecting the independent variables ‘Satisfaction’ and ‘Trust after the Interaction.’

In order to answer the fourth research question regarding the effect of ‘Satisfaction’ on ‘Trust after the Interaction’ and vice versa, both variables were treated as an independent as well as a dependent variable.

To answer the last research question concerning the effect ‘Satisfaction’ and ‘Trust after the Interaction,’ respectively, have on the ‘Usage Continuance Intention,’ the latter variable was treated as a dependent, while Satisfaction and Trust after the Interaction were treated as independent variables.

3. Results

The following section will be divided into confirmatory factor analysis, the correlation analysis between the CUS and the UMUX-LITE, and ends with the regression analyses. The R script can be found in Appendix D.

3.1 Confirmatory factor analysis

The assumptions of normality of the data measured by the Shapiro-Wilk test showed significant nonnormality (p_{norm}'s< 0.001). Hence, the robust maximum likelihood (MLR) method was used for each model modification. For the following, all scores gathered for each model can be found in Table 2.

The initial factor model M1 showed moderate scores along the measurement variables: $\chi^2_{M1} = 314.409$ with $p_{\chi 2} < 0.001$; $RSMEA_{M1} = 0.084$; $SRMR_{M1} = 0.048$; and $CFI_{M1} = 0.926$. However, the first item (CUS_1) showed a negative variance estimate ($var_{CUS\_1} = -0.058$). Consequently, the researchers decided to remove CUS_1 from the model.

In the second model (M2), with the removed item CUS_1, all indicators showed significant positive factor loadings and standardized coefficients ranging from 0.029 to 0.052 (p's< 0.001). However, fit indices for the model varied, representing a rather poor fit to the data: $\chi^2_{M2} = 298.524$ with $p_{\chi 2} < 0.001$; $RSMEA_{M2} = 0.09$; $SRMR_{M2} = 0.048$; and $CFI_{M2} = 0.919$.

A review of residual correlations and modification indices should determine whether including additional parameters in the model may improve model fit. The largest modification index (mi = 36.459) indicated that the model would be improved if the error terms of item CUS_11 ("The chatbot gives me the appropriate amount of information") and item CUS_12 ("The chatbot only gives me the information I need") would be permitted to covary. This was also consistent with the observation of a large residual correlation (coefficient = 1.719) between these variables. Theory suggests that a high modification index may imply that the structure of the model might have not correctly been captured with the current model but that another factor might be suitable (Moosbrugger & Kelava, 2020). To put it differently, items that show a covariance with each other might measure something different than the other items of that factor and are allowed to be put in a separate factor (Barney et al., 2021). In order to test for this, both items were taken into further evaluation and were compared with the other items of the factor F3 (CUS_10: "I find that the chatbot understands what I want and helps me achieve my goal"; CUS_13: "I feel like the chatbot's responses were accurate"). First, the residual correlations of all four items were compared with each other. Hereby, no other than CUS_11 and CUS_12 showed a high correlation, indicating that only those two might measure something similarly (see Appendix D1). Next, the modification indices were investigated once more. However, no other covariance between the items CUS_10 to CUS_13 were found except the already indicated one between CUS_11 and CUS_12. Consequently, this was seen as another indicator that no other item is measuring the same underlying idea of items 11 and 12. Lastly, the meaning of the four items of factor F3 were evaluated on a common sense basis. Hereby, it was found that both items CUS_11 and CUS_12 describe the same underlying purpose, namely the information the chatbot gives the user, while this was not the case for any other of the items of factor F3. Based on these insights it was decided to create a new factor (F4) with the variables CUS_11 and CUS_12 to improve model fit in a subsequent model M3.

Results of M3 showed an increase in model fit, however, the cut off score was not achieved (see Table 2). Therefore, factor loadings, residual correlations and modification indices were reviewed to see which modification might increase model fit. Hereby, item CUS_4 ("I was immediately made aware of what information the chatbot can give me") showed the lowest factor loading with 0.620 which is why further investigation was done with this item. First, the modification indices were evaluated which revealed that CUS_4 was suggested to covary with item CUS_15 ("My waiting time for a response from the chatbot was short;" mi = 13.517). This covariation was not found to be fitting as, by evaluating the items with common sense, CUS_4 and CUS_15 are intended to measure something completely different. However, the modification indices also suggested to let CUS_4 covary with the factors F1, F5, and F6, which might include that the item CUS_4 is measuring something which is not fully captured by its current factor F2, but not fully captured by any other factor either. Hence, an investigation of the normalized residual variance-covariance matrix should reveal with which items CUS_4 is correlating. Hereby, high correlations were found for CUS_4 with CUS_2 (coefficient = 2.398), with CUS_12 (coefficient = 1.002), with CUS_14 (coefficient = 2.339), and with CUS_15 (coefficient = 2.353). However, no such high correlation was found with any other item of its initial factor F2 (see Appendix D1). This was seen as another indicator that the item CUS_4 might measure something which is not fully captured by the model or any other factor. Due to that fact, it was decided to drop CUS_4 in the subsequent model M4 to improve model fit.

A re-ran of the analysis with M4 led to an increase in model fit but the cut off scores were still not achieved (see Table 2): $\chi^2_{M4} = 219.434$ with $p_{\chi 2} < 0.001$; $RSMEA_{M4} = 0.089$; $SRMR_{M4} = 0.044$; and $CFI_{M4} = 0.936$. Thus, another review of the data was conducted. First, it was seen that the item CUS_7 (“The chatbot was able to make references to the website or service when appropriate”) has the lowest factor loading of 0.619. An evaluation of the modification indices only suggested a covariation of CUS_7 with CUS_9 (“The chatbot’s responses were easy to understand”). However, comparing the items with each other using common sense, it was no underlying structure found which both of these items would share, as referencing to the website and easy responses were not seen to be similar. In the end, it was decided to keep item CUS_7 anyway as a factor loading of 0.619 is still sufficient (Peterson, 2000). Furthermore, this item showed high residual correlations with other items of factor F2, suggesting that their underlying messages are related with each other (see Appendix D1). Moreover, it was argued that this item might measure something which is important in the context of user-chatbot interactions in customer service as chatbots’ capabilities are not always sufficient for the user (Kvale et al., 2021).

As a second step, the focus then shifted to other modification indices that were suggested. Hereby, CUS_3 (“Communicating with the chatbot was clear”) and CUS_9 were suggested to covary (mi = 21.926). Further, a covariation of CUS_6 (“The chatbot was able to keep track of context”) with CUS_8 (“The chatbot could handle situations in which the line of conversation was not clear”) was suggested (mi = 21.149), as well as CUS_5 (“The interaction with the chatbot felt like an ongoing conversation”) with CUS_6. Hence, it seemed that there was a lot of correlation between the items of factor F2 which was, indeed, confirmed by the residual correlation table (see Appendix D1). This suggested, that splitting up these items into new factors (as done in model M3) would not be supported in any way. Nevertheless, a review of the items with keeping in mind the residual variance-covariance matrix let the researchers suggest that a covariation of the above mentioned items would be supported: if the communication was clear (CUS_3), then the chatbot’s responses should be easy to understand (CUS_9) in a similar way. Moreover, if the chatbot is able to handle situations in which the line of conversation was not clear (CUS_8), then it should automatically be suggested that the users have the feeling that the chatbot can keep track of context (CUS_6). Lastly, if the chatbot can keep track of context (CUS_6), then it should also be felt like an ongoing conversation (CUS_5) by the users. To conclude, it was decided to let the above mentioned items covary with each other in the following model M5 to improve model fit.

As reported in Table 2, the analysis performed on M5 led to a better fit as follows: $\chi^2_{M5} = 177.977$, $p_{\chi2} < 0.001$; $RMSEA_{M5} = 0.080$ (≤ 0.080); $SRMR_{M5} = 0.042$ (< 0.080); and $CFI_{M4} = 0.951$ (> 0.95). AIC and BIC also decreased compared to the previous models.

Based on these improvements in terms of fitness, M5 was selected as the best solution for the factorial structure. Overall, the scale in line with M5 indicates a high reliability (α = 0.92). A visual representation of M5 can be found in Appendix D1. The overview of the solution suggested by M5 for the CUS can be found in Appendix E. The new scale is composed by 13 items and six components, with one new factor: “Perceived information representation” (F4).

Table 2

*Fit indices of the initial model (M1) and modified models (M2-M5) from the confirmatory factor analysis of the CUS (Borsci et al., Under Review). The specific fit indices are Chi-square goodness of fit statistics (χ2), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). ***p< 0.001*

	M1	**M2**	**M3**	**M4**	**M5**
χ^2 (p> 0.05)	314.409***	298.524***	267.950***	219.434***	177.997***
RSMEA (≤ 0.08)	0.084***	0.090***	0.088***	0.089***	0.080***
SRMR (< 0.08)	0.048	0.048	0.047	0.044	0.042
CFI (>0.95)	0.926	0.919	0.928	0.936	0.951
AIC	15822.855	14975.924	14943.646	13788.544	13745.101
BIC	16034.402	15171.506	15159.185	13992.108	13960.641

3.2 Correlation between the UMUX-LITE and the 15- and 13-items CUS

To answer the second research question and to assess the CUS' concurrent validity, the correlation between the original 15-items CUS and the UMUX-LITE as well as the modified 13-items CUS was examined. Both CUS versions showed a strong correlation with the UMUX-LITE. When investigating the correlation with each factor in specific, the UMUX-LITE had a good correlation with factor F2 of each the 15- and 13-items CUS, and a moderate correlation with factor F3 for the 15-items CUS. However, for the 13-items CUS a strong correlation of the UMUX-LITE with F3 could be found. Moreover, the correlations with factor F4 of the modified model were good. All other correlations were found to be weak (see Table 3).

Table 3

*Correlations measured with Spearman's rank-order correlation between the UMUX-LITE and the 15- and 13-items CUS questionnaire, respectively. ***p< 0.001*

	UMUX-LITE
15-items CUS	0.804***
(F1) Perceived accessibility to chatbot functions	0.356***
(F2) Perceived quality of chatbot functions	0.770***
(F3) Perceived quality of conversation and information provided	0.649***
(F4) Perceived privacy and security	0.213***
(F5) Time response	0.490***
Modified 13-items CUS	0.813***
(F1) Perceived accessibility to chatbot functions	0.269***
(F2) Perceived quality of chatbot functions	0.763***
(F3) Perceived quality of conversation and information provided	0.808***
(F4) Perceived information representation	0.726***
(F5) Perceived privacy and security	0.213***
(F6) Time response	0.489***

3.3 Regression analyses

For answering the research questions three, four, and five, the relationships of the scales Initial trust, Satisfaction, Trust after the interaction, and Usage continuance intention, respectively, were investigated. An overview of the significant effects are summarized in Table 4, and a visual representation of them can be found in Appendix D2.

Concerning the third research question, four analyses were run, with 'Disposition to Trust' x 'Satisfaction,' 'Personal Innovativeness' x 'Satisfaction,' 'Disposition to Trust' x 'Trust after the Interaction,' and 'Personal Innovativeness' x 'Trust after the Interaction.' Results showed only a significant effect of 'Personal Innovativeness' on 'Satisfaction' (b_{PiS} =

0.116, $t_{PiS}(389) = 2.530$, $p_{PiS} = 0.012$, $R^2_{PiS} = 0.170$). Neither of the other analyses showed a significant relationship with each other.

In order to answer the fourth research question, analyses were performed for ‘Satisfaction’ x ‘Trust after the Interaction’, and ‘Trust after the Interaction’ x ‘Satisfaction’ respectively. For both analyses, significant regression equations were found ($b_{SaTr} = 0.892$, $t_{SaTr}(389) = 38.717$, $p_{SaTr} < 0.001$, $R^2_{SaTr} = 0.807$; $b_{SaTr} = 0.884$, $t_{TrSa}(389) = 38.473$, $p_{SaTr} < 0.001$, $R^2_{TrSa} = 0.807$).

For the last research question two regressions were run with the parameters ‘Usage Continuance Intention’ x ‘Satisfaction,’ and ‘Usage Continuance Intention’ x ‘Trust after the Interaction.’ Both results revealed a positive significant relationship ($b_{UciSa} = 0.723$, $t_{UciSa}(389) = 20.859$, $p_{UciSa} < 0.001$, $R^2_{UcoSa} = 0.522$; $b_{UciTr} = 0.780$, $t_{UciTr}(389) = 24.867$, $p_{UciTr} < 0.001$, $R^2_{UcoTr} = 0.608$).

Table 4

*Significant regression correlations of the variables Personal innovativeness, Satisfaction, Trust after the interaction, and Usage continuance intention with their estimated values, standard errors (Std.error), and t-statistics including the degrees of freedom (df). *p< 0.05; **p< 0.01; ***p< 0.001*

	Value	**Std.error**	**t-statistics (df)**	**R^2**
Personal innovativeness x Satisfaction	0.116	0.046	2.530 (389)*	0.170
Satisfaction x Trust after the interaction	0.892	0.023	38.717 (389)***	0.807
Trust after the interaction x Satisfaction	0.884	0.023	38.473 (389)***	0.807
Usage continuance intention x Satisfaction	0.722	0.035	20.859 (389)***	0.522
Usage continuance intention x Trust after the interaction	0.780	0.031	24.867 (389)***	0.608

4. Discussion

This study aimed at testing the newly developed CUS, a scale for chatbot-satisfaction investigation, developed by Borsci et al. (Under Review), to see if the factor structure could be confirmed. Additionally, a correlation analysis was done between the CUS and the UMUX-LITE for external validation of the CUS. Lastly, regression analyses were done to determine the relationship between trust, satisfaction, and usage continuance intention. Hereby, the effects of 'Disposition to Trust' and 'Personal Innovativeness' (following summarized as *initial trust*) on 'Satisfaction' and 'Trust after the Interaction' were investigated. Further, the effect of 'Satisfaction' on 'Trust after the Interaction' and vice versa was estimated. Lastly, it was analyzed if 'Usage Continuance Intention' is affected by 'Satisfaction' and 'Trust after the Interaction.' In detail, five different research questions were examined which will be evaluated in the following.

4.1 Psychometric properties of CUS

The first research question was: "Can the psychometric properties of the CUS be confirmed with a factor loading on five factors and a reliability over 0.7?" The results could not confirm the five-factor structure of the CUS and the model showed a poor fit with the sample. However, reliability of the subsequent modified model was high.

First, the data were tested for normal distribution which hypotheses had to be rejected, which implies the data was not normally distributed. Although this is a sign that the model data is not good, the assumption of normality does rarely met with empirical data (Benson & Fleishman, 1994). Nonetheless, the robust maximum likelihood (MLR) method was used for each model, as this statistically corrects standard errors and chi-square test statistics and, thereby, enhances the robustness against departures from normality (Li, 2016).

In the initial model of analysis, a negative variance for item 1 was detected, which is known as a "Heywood case" (Harman & Fukuda, 1966). This means that the factor solution might reflect the observed correlations perfectly but lacks the basic requirement to be between 0 and 1 (Harman & Fukuda, 1966). This is a sign of too much collinearity. There are several reasons for a Heywood case to arise, one of which is sampling fluctuations (Kolenikov & Bollen, 2012). As the given sample is rather small (Benson & Fleishman, 1994) and the data gathering was rather inadequate due to several circumstances (see section 4.4), this might be a reason why the model shows poor fit with this sample.

In order to fit the model as best as possible, a modification of the model based on the factor loadings, examination of modification indices, and normalized residual variance-covariance matrices were done. Based on the examination of these values, it was suggested to include one new factor, namely “Perceived information representation” (F4) with items 11 and 12. Furthermore, one item was dropped (CUS_4), and three covariances were included in the final model (CUS_3~~CUS_9; CUS_5~~CUS_6; CUS_6~~CUS_8).

When reviewing items 11 and 12 prior to post-hoc modification, it was noted that items 11 and 12 were the only items that specifically addressed the quantity of information the chatbot would give the user. This might indicate multidimensionality within the latent construct of the factor ‘Perceived quality of conversation and information provided’ (Barney et al., 2021). Hence, this might imply that model fit would be increased by modelling these items in separate distinct factors if those are the only items that measure this underlying construct (Barney et al., 2021; Moosbrugger & Kelava, 2020). As this was seen to be provable by comparison with the other items’ meaning, and by investigation of the normalized residual variance-covariance matrix and modification indices, it was decided to follow this suggestion, and creating a new factor F4. This factor was subsequently named “Perceived information representation” while the original factor F3 was shortened to “Perceived quality of conversation.”

In a next step, the model M3 was evaluated and it was decided to drop item CUS_4 from the scale as it seemed not to be captured by the model. A reason why this might have happened may lie in the way the participants have interacted with the chatbots during the study. As there were clear goals or tasks, respectively, many participants were observed to type in their request right away without reading through the first messages of the chatbots in which the capabilities of the chatbot were explained in most cases. Another reason might lie in the wording of the item, as it might not have been clear to the participants what it exactly implies. Consequently, further research could investigate if a rewording of the item, such as asking “I was immediately made aware of the capabilities of the chatbot regarding information provision” instead of the current wording, might have a positive effect. In the end, it is not clear why CUS_4 showed to not fit with the current sample which is why further investigations on this item might be advisable.

Lastly, covariances of CUS_3 with CUS_9, CUS_5 with CUS_6, and CUS_6 with CUS_8 were allowed as those items were part of the same factor, and modification indices and variance-covariance values were supporting this decision. Hereby, after investigating and comparing the indicated item pairs with each other, each of these pairs showed to be connected in their underlying message. Nonetheless, it was not doable to split these item pairs up into separate factors, like it was done before; all items of the factor clearly showed further correlations with each other. Explanations for this might be that all of these items still displayed the ‘Perceived quality of the chatbot functions,’ just as it was indicated by the factor’s name. When looking at the different items it can be seen that all of them are implying the communication or responses of the chatbot, which are aspects that are related with each other. Hence, they are all using the same underlying message. Nevertheless, it is still possible to find either the one or the other in the statements which make these covariation possible as well. Eventually, this might just be another indicator that the model showed poor fit with the sample. Accordingly, the subsequent and final model M5 may not be suitable for other samples as well and further investigations should be done here.

Although this last model (M5) fulfilled the cutoff scores set beforehand, this model should be used with caution. As already mentioned, the sample was not a good fit for the model as the normality of the data was not given, and in the beginning, a Heywood case could be detected. Hence, even the eventual fulfillment of cutoff scores might not exclude sample effect, thus, this model might not be replicable in other studies (Moosbrugger & Kelava, 2020). Moreover, the Chi-square statistics of fit was still high and significant for the last model, similarly as the AIC and BIC. Literature suggests that a smaller value of the AIC and BIC imply a better model fit, but it does not define specific which scores are better or worse (Wang & Liu, 2006). However, compared to other studies that used confirmatory factor analysis, the scores obtained in this study remained high. Summing it up, this is another aspect of why this model might not capture what the CUS is intending to measure. Consequently, subsequent studies should be consulted before clearly interpreting the model M5 in a specific direction and comparing it with the initially found psychometric properties of the CUS.

As a last aspect, the revised model M5 showed high reliability with a Cronbach’s alpha higher than 0.9. As this score can be seen to be very good, it should also be seen with caution as it might indicate redundancy in items (Taber, 2018). Therefore, the result of Cronbach’s alpha might be another argument for using the modified model M5 with caution.

4.2 Comparing the 14- and 15-items CUS with the UMUX-LITE

The second research question was “Does the 15-items CUS correlate with the UMUX-LITE?”. The results of the Spearman rank-order correlation test showed that the UMUX-LITE had a strong relation to both the 15- and 13-items CUS which indicates that all three scales are measuring the same underlying construct. Moreover, a strong correlation between the CUS’ ‘Perceived quality of chatbot functions’ factor (F2) and the UMUX-Lite, as well as a moderate correlation between ‘Perceived quality of conversation and information provided’ (F3) of the 15-items CUS and the UMUX-LITE could be detected. For the modified 13-item model of the CUS, the factor ‘Perceived quality of conversation’ (F3) showed a strong correlation, while the factor ‘Perceived information representation’ (F4) showed a good correlation with the UMUX-LITE. However, the correlations between the UMUX-LITE and the remaining factors ‘Perceived accessibility to chatbot functions’ (F1), ‘Perceived privacy and security’ (F4/F5), and ‘Time response’ (F5/F6) were weak to very weak for both models. This suggests that these latter factors measure different aspects of user satisfaction than the UMUX-LITE does.

The finding that the UMUX-LITE is not reflected in all factors of the user experience in chatbots, is in line with previous findings by Tariverdiyeva and Borsci (2019), and Silderhuis and Borsci (2020). Silderhuis and Borsci (2020) also found a strong overall correlation between the UMUX-LITE and their USIC, but for the specific factors, only the ‘Communication quality’ factor was found to have a strong relation to the UMUX-LITE. All other factors showed weak to very weak correlations. Moreover, Tariverdiyeva and Borci (2020) concluded in their research that the UMUX-LITE might be able to inform about the usability of the chatbot but that it is missing relevant aspects to explain the whole user experience with a chatbot interface. Hence, these consistent findings suggest that the concept of user satisfaction is different in the UMUX-LITE than in the CUS, and the former might only reflect segments of it.

Arguably, reasons for the weak correlations between the UMUX-LITE and the factors ‘Perceived accessibility to chatbot functions’ (F1), ‘Perceived privacy and security’ (F4/F5), and ‘Time response’ (F5/F6) may lie in the diagnostic character of the CUS. The CUS is designed on a more complex construct and means to provide a more complete picture of user satisfaction with chatbots. Thereby, it is covering several aspects found to be important in literature and different analyses, and is, moreover, explicitly designed for chatbots. The UMUX-LITE, however, is designed for system interfaces that are not as complex as chatbots (Følstad et al., 2018; van den Bos & Borsci, 2019). Therefore, it is reasonable to assume that the CUS provides a more elaborative view on user satisfaction than the UMUX-LITE does. Therefore, the factors ‘Perceived accessibility to chatbot functions’ (F1), ‘Perceived privacy and security’ (F4/F5), and ‘Time response’ (F5/F6) are seen to reflect valuable support for the CUS’ diagnostic criteria and, therefore, should be held.

4.3 Relationships between initial trust, satisfaction, trust after the interaction, and usage continuance intention

The third research question was "Do disposition to trust and personal innovativeness have an effect on satisfaction and trust after the interaction with the chatbot?" The analyses showed a significant effect only for the relationship between 'Personal Innovativeness' and 'Satisfaction.' Hence, it can be concluded that personal innovativeness affects the perception of satisfaction, although this influence was not very strong. Consequently, the results were not in line with previous research by McKnight et al. (2002) or Gefen and Straub (2004), where disposition to trust was found to have an (at least indirect) effect on trust formation. One reason for this might be that McKnight et al. (2002) divided initial trust into many more subconstructs than just disposition to trust. As the constrain on the participants should be as low as possible in this study, it was decided to only include disposition to trust as an initial trust measure as it had an at least indirect effect on trust formation in the initial study (McKnight et al., 2002). Nevertheless, the fact that McKnight et al. (2002) found it to be indirectly affecting trust could be the reason for showing no effect on satisfaction or trust after the interaction in this sample. Therefore, further investigations should be done if a combination of several initial trust constructs would lead to significant effects. Another reason why disposition to trust might not have been significant is that the questions were formulated regarding human-human relationships but not towards human-technology or human-chatbot relationship. In the examples of McKnight et al. (2002) and Gefen and Staub (2004), this concept might have been suited very well as they were investigating initial trust towards eCommerce. In eCommerce, the user has a greater feeling of social presence even though they are interacting through a website (Gefen & Straub, 2004). In the context of chatbots, however, people might lose this feeling as they know that they are talking to, and evaluating something nonhuman (Jenkins et al., 2007). Concluding, there could be many reasons why disposition to trust has not shown a significant effect on satisfaction or trust after the interaction in this study, which might be a topic of further investigation. Lastly, personal innovativeness may have found to have a positive significant effect on satisfaction as early adopters, i.e. those people who are willing to try out new products at a very early stage of product implementation, also have a more positive attitude towards this new product or technology (Nordheim et al., 2019). Since chatbots are only getting more popular by now (Dale, 2016), the people who are open to exploring new technology might, consequently, also be more open towards this new upcoming trend of chatbots. In summary, early adopters have a more positive mindset towards chatbots and, hence, are more easily satisfied with them.

This might explain why in this research a relationship between ‘Personal Innovativeness’ and ‘Satisfaction’ was found.

Next, the fourth research question was “Does satisfaction with chatbots have an effect on trust after the interaction and vice versa?” The results of the study showed that both variables are influencing each other. This is not in line with the findings of Lankton et al. (2014), who could only find an influence of satisfaction on trust but not the other way around. On the one hand, the effect of satisfaction on trust after the interaction in context of chatbot usage is furtherly confirmed by Følstad et al. (2018), who came up with an initial model about trust in chatbots for customer service. In their model, they distinguish between chatbot-related factors, such as ‘Interpretation and advice’ or ‘Professional appearance,’ and factors concerning the service environment. Especially the former is, thereby, showing that similar aspects, which are implemented in the CUS to measure for satisfaction, are influencing the building of trust towards chatbots. Therefore, this suggests that a high satisfaction level is predicting a higher level of trust. Nevertheless, Balaji and Borsci (2020) also included items of trust in their chatbot satisfaction scale as they found that trust is an indicator of satisfaction. Hence, all this might be an indication that ‘Satisfaction’ and ‘Trust’ are influencing each other, as high satisfaction might suggest a higher trust level, while trust is incorporated in a higher satisfaction perception. Thus, the difference to Lankton et al.’s (2014) study might, again, lie in the different technologies that were used back then compared to the one used in this study. Conclusively, this might explain why there was an effect towards both ‘Trust after the Interaction’ on ‘Satisfaction,’ and ‘Satisfaction’ on ‘Trust after the Interaction,’ which was not found in the initial study by Lankton et al. (2014).

Lastly, the fifth research question was “Is usage continuation intention affected by trust after the interaction and satisfaction measured by the CUS?” To answer this question, a regression analysis was conducted whereof the results showed that both concepts have a moderate effect on usage continuance intention. However, the effect of ‘Trust after the Interaction’ was higher than that of ‘Satisfaction.’ Hence, it can be concluded that trust after the interaction has a better effect on usage continuation intention than satisfaction does. This finding is in line with the one of Lankton et al. (2014) for which trust also displayed a higher effect on usage continuance intention than satisfaction did. Nevertheless, the differences found in this study are not very large. Conclusively, it can be said that trust is at least as important as satisfaction for the implementation of devices by users but satisfaction has a good effect on it as well. Therefore, both concepts should be kept in mind when designing and implementing a chatbot.

4.4 Limitations and future research

The first limitation of the current study is concerning the sample that was used. Due to time constraints, snowball sampling was used as a sampling method instead of probability sampling methods. A disadvantage of this sampling method is that it generates biases in the sampling data, as the sampling is merely done through the social connections of the participants (Etikan et al., 2018). As a result, the sampling population has similar characteristics which has an effect on the gathered data. Furthermore, the sample size was at the lower cutoff for executing a confirmatory factor analysis with valid results. Sample size has a significant effect on model fit and misspecifications in confirmatory factor analysis (Shevlin & Miles, 1998). Therefore, the sample size might also be a predictor of the poor model fit for the sampling data. Consequently, future research should investigate the psychometric properties of the CUS with larger and more varied samples of individuals.

A second limitation of the research is that the participant only interacted with the chatbot during one task. Hereby, the duration of interaction with the chatbot varied significantly: some chatbots only needed interaction of less than thirty seconds while other interactions lasted for several minutes. Consequently, some questions asked by the CUS were difficult to answer for the participants because the topics that were asked did not occur during the interaction. Hence, participants had to answer these questions with just their intuition which displays another limitation in the validity of the study. Therefore, future research should assess the tasks by trying to give the participants a fuller experience of the chatbots' capabilities. This suggestion is also supported by research which found that the opinion about a system changes over a period of usage (Borsci et al., 2015; Tsakonas & Papatheodorou, 2008). Conclusively, making tasks longer lasting for a better chatbot experience and assessing how the CUS is operating over a period of usage might be something future research could investigate.

Another limitation arose due to the Covid-19 pandemic, because of which the study had to be done online. Even though participants were told to be in a quiet closed up room, disturbances could not fully be minimized. As participants participated from home, disturbances through outside noise, through people accidentally entering the room or talking to the participants from outside caused temporary distractions from the tasks. Hence, it was difficult to create a controlled environment for the study. Furthermore, as each participant used a different computer for participation, chatbots were not displayed similarly for all participants and were sometimes even lagging; during a few sessions chatbots would load only after waiting for it unusually long which displeased the participants. In the end, all this might have had an effect on the outcome found in this study. Consequently, future research might try testing the CUS in a more controlled environment to minimize influences on the results from outside.

Lastly, the questions used in this study testing for initial trust, trust after the chatbot interaction, and usage continuance intention were developed for technologies, but not for chatbots in specific. Therefore, some (formulations of the) questions were often not clear to the participant as they asked for something the chatbot did not display (e.g. questions about a help function). This, subsequently, caused confusion by the participant and might have had an effect on the results of this study. Følstad et al. (2018) already developed an initial model about aspects that affect trust in chatbots, however, it is just a theoretical basement so far. Future research could investigate questions for reliably testing for trust in and towards chatbots in specific.

5. Conclusion

The current study investigated the psychometric properties of the newly developed Chatbot Usability Scale (CUS). The results could not confirm the five-factor structure of the initial CUS but suggested an enlargement of factors into a six-factor model while maintaining high reliability. Additionally, correlations between the 15- and 13-items CUS and the UMUX-LITE were estimated, which indicated that all three scales are measuring the same underlying construct. Thirdly, this study investigated the relationships of trust, satisfaction, and usage continuance intention. The effects of initial trust on ‘Satisfaction’ and ‘Trust after the Interaction’ were examined, whereby only ‘Personal Innovativeness’ significantly correlated with satisfaction. Furthermore, the relationship between ‘Satisfaction’ and ‘Trust after the Interaction’ was investigated. Hereby, both trust after the interaction and satisfaction were detected to be dependent on each other. Lastly, ‘Satisfaction’ and ‘Trust after the Interaction’ and their effects on ‘Usage Continuance Intention’ were investigated. It was found that both are indicators for usage continuance intention. Overall, the study showed a poor fit between the CUS and the sample, which might have been influenced by sampling method and sample size. Moreover, the tasks used could be improved so that participants get better impressions of the capabilities of the chatbot they are evaluating. Nevertheless, the CUS might be a good starting point for evaluating satisfaction in chatbots which justify further research on it.

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. https://doi.org/10.1016/j.chb.2018.03.051

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

Barney, J. L., Barrett, T. S., Lensegrav-Benson, T., Quakenbush, B., & Twohig, M. P. (2021). Confirmatory factor analysis and measurement invariance of the Cognitive Fusion Questionnaire-Body Image in a clinical eating disorder sample. *Body Image, 38*, 262-269. https://doi.org/10.1016/j.bodyim.2021.04.012

Benson J., & Fleishman, J. A. (1994). The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. *Quality & Quantity, 28*, 117-136. Retrieved from https://link.springer.com/content/pdf/10.1007/BF01102757.pdf

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. https://doi.org/10.1080/10447318.2015.1064648

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

Brandtzæg, P. B., & Følstad, A. (2017). Why people use chatbots. In *INSCI 2017: Internet Science*, 377-392. https://doi.org/10.1007/978-3-319-70284-1_30

Dale, R. (2016). Industry Watch: The return of the chatbots. *Natural Language Engineering 22*(5), 811–817. doi:10.1017/S1351324916000243

Epskamp, S., Stuber, S., Nak, J., Veenman, M., & Jorgensen, T. D. (2019). semPlot: Path diagrams and visual analysis of various SEM packages' output. Retrieved from https://cran.r-project.org/web//packages/semPlot/semPlot.pdf

Etikan, I., Alkassim, R., & Abubakar, S. (2015). Comparision of Snowball Sampling and

Sequential Sampling Technique. *Biometrics & Biostatistics International Journal, 3*(1), 00055. DOI: 10.15406/bbij.2015.03.00055

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers, 22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

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

Følstad, A., & Brandtzæg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience 5*(3), 1-14. https://doi.org/10.1007/s41233-020-00033-2

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In *International Conference on Internet Science*, 194-208. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-01437-7_16

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1-9. https://doi.org/10.1145/3342775.3342784

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. *Omega, 32*, 407 – 424. doi:10.1016/j.omega.2004.01.006

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems (ICIS)*, 1-13. Retrieved from https://www.researchgate.net/profile/Ulrich-Gnewuch/publication/320015931_Towards_Designing_Cooperative_and_Social_Conversational_Agents_for_Customer_Service/links/59c8d1220f7e9bd2c01a38a5/Towards-Designing-Cooperative-and-Social-Conversational-Agents-for-Customer-Service.pdf

Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro-Wilk test with known mean. *REVSTAT 14*(1), 89–100. Retrieved from https://www.ine.pt/revstat/autores/pdf/rs160105.pdf

Harman, H. H., & Fukuda, Y. (1966). Resolution of the Heywood case in the minres solution. *PSYCHOMETRIKA, 31*(4), 563-571. Retrieved from https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/content/pdf/10.1007/BF02289525.pdf&casa_token=HVNJn8ssrDIAAAAA:E7bY7_SxKVdA

KiUPhCMnGhuapOuhWxZQMQsF4iLNe7IcBvE12PG4BryclQgGu8sMMGLXp9HvPrbPXxBuxA

Jenkins, M.-C., Churchill, R., Cox, S., & Smith, D. (2007). Analysis of user interaction with service oriented chatbot systems. *Proceedings International Conference of Human-Computer Interaction*, 76–83. https://doi.org/10.1007/978-3-540-73110-8

Kolenikov, S., & Bollen, K.A. (2012). Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?. *Sociological Methods & Research* *41*(1) 124–167. DOI: 10.1177/0049124112442138

Kvale, K., Freddi, E., Hodnebrog, S., Sell, O. A., & Følstad, A. (2021). Understanding the user Experience of Customer Service Chatbots: What Can We Learn from Customer Satisfaction Surveys?. In: *CONVERSATIONS 2020, LNCS 12604*, 205–218. https://doi.org/10.1007/978-3-030-68288-0_14

Lankton, N., McKnight, D. H., & Thatcher, J. B. (2014). Incorporating trust-in-technology into Expectation Disconfirmation Theory. *Journal of Strategic Information Systems, 23*, 128–145. http://dx.doi.org/10.1016/j.jsis.2013.09.001

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

Li, C. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res, 48*, 936–949. DOI 10.3758/s13428-015-0619-7

McKnight, D. H., Carter, M., & Clay, P. (2009). Trust in technology: development of asset of constructs and measures. In *DIGIT 2009 Proceedings. 10*, 2-12. Retrieved from http://aisel.aisnet.org/digit2009/10

McKnight, D. H., & Chervany, N. L. (1996). The meaning of trust. *University of Minnesota,* 1-87. Retrieved from https://www.researchgate.net/publication/239538703_The_Meanings_of_Trust

McKnight, D. H., Choudhurry, V., & Kacmar, C. (2002). Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research, 13*(3), 334–359. Retrieved from https://www.researchgate.net/publication/220079434_Developing_and_Validating_Trust_Measures_for_e-Commerce_An_Integrative_Typology

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. Retrieved from https://www.jstor.org/stable/259290

Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* (3rd ed.). Springer. https://doi.org/10.1007/978-3-662-61532-4

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72-78. Retrieved from https://dl.acm.org/doi/pdf/10.1145/191666.191703?casa_token=bzWXIgDS60QAAAAA:_Xc3BrSsAZTVIiP3uI_xC9dH9QpZVeVFPngTYqWiJry3foZ5ne516Myo_v2Nmz4WVj0e8w89Cg_qYw

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An Initial Model of Trust in Chatbots for Customer Service - Findings from a Questionnaire Study. *Interacting with Computers, 31*(3), 317-335. doi: 10.1093/iwc/iwz022

Nuruzzaman, M., & Hussain, O. K. (2018). A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In *IEEE 15th International Conference on e-Business Engineering (ICEBE)*, 54-61. DOI 10.1109/ICEBE.2018.00019

Peterson, R. A. (2000). A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters 11*(3), 261-275. Retrieved from https://link.springer.com/content/pdf/10.1023/A:1008191211004.pdf

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK authors, Heisterkamp, S., Van Willingen, B., Ranke, J., & R-core. nlme: Linear and nonlinear mixed effects models (Version 3.1-152). Retrieved from https://cran.r-project.org/web/packages/nlme/nlme.pdf

Revelle, W. (2021). psych: Procedures for personality and psychological research. (Version 2.1.6). Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: a cross discipline view of trust. *Academy of Management Review, 23*(3), 393-404. Retrieved from https://www.jstor.org/stable/259285

Rosseel, Y, Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., … Du, H. (2021). lavaan: Latent variable analysis (Version 0.6-8). Retrieved from https://cran.r-project.org/web//packages/lavaan/lavaan.pdf

Shevlin, M., & Miles, J. N. V. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences, 25*, 85-90. https://doi.org/10.1016/S0191-8869(98)00055-5

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

Skjuve, M. B., & Brandtzaeg, P. B. (2019). Measuring user experience in chatbots: An approach to interpersonal communication competence. In *International Conference on Internet Science*, 113–120. https://doi.org/10.1007/978-3-030-17705-8_10

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res Sci Educ, 48*, 1273–1296. DOI 10.1007/s11165-016-9602-2

Tariverdiyeva, G., & Borsci, S. (2019). Chatbots' perceived usability in information retrieval tasks: An exploratory analysis. [Master Thesis]. University of Twente, Enschede, The Netherlands. http://essay.utwente.nl/77182/

Tsakona, G., & Papatheodorou, C. (2008). Exploring usefulness and usability in the evaluation of open access digital libraries. *Information Processing and Management, 44*, 1234–1250. doi:10.1016/j.ipm.2007.07.008

Van den Bos, M., & Borsci, S. (2021). Testing a scale for perceived usability and user satisfaction in chatbots: Testing the BotScale. [Master Thesis]. University of Twente, Enschede, The Netherlands.

Van der Goot, M. J., Hafkamp, L., & Dankfort, Z. (2021). Customer Service Chatbots: A Qualitative Interview Study into the Communication Journey of Customers. In: *CONVERSATIONS 2020,* 190–204. https://doi.org/10.1007/978-3-030-68288-0_13

Wang, Y. D., & Emurian, H. H. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior, 21*, 105-125. https://doi.org/10.1016/j.chb.2003.11.008

Wang, Y., & Liu, Q. (2006). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research 77*, 220–225. doi:10.1016/j.fishres.2005.08.011

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C. … Dunnington, D. (2021). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (Version 3.3.4). Retrieved from https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *NATURE GENETICS, 46*(2), 100-106. doi:10.1038/ng.2876

Appendix

Appendix A: Research information sheet and informed consent

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

Purpose of the research

This research aims to retest a scale, with which chatbots can be evaluated. A chatbot is a program with which you can chat trough text and it will give you answers based on what you say, for example used in customer service. Additionally, we will test for trust pre and post the interaction with the chatbot.

Study content

You will receive some tasks from me, interact with a few chatbots and after this interaction with each chatbot you will have to fill out a scale to evaluate the chatbot and an additional existing scale to compare our scale with. Lastly, questions regarding different areas of trust have to be filled out. The study will take around an hour to 75 minutes, and there are no risks attached to your participation.

Data acquisition

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

Contact

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

	Select your answer	
	Yes (1)	No (2)
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. (1)	○	○
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. (2)	○	○
I understand that taking part in the study involves answering questions about my demographics, performing tasks and interacting with chatbots online, filling out three scales about each of the ten chatbots I have interacted with online, and have an online call with the researcher which is being recorded. (3)	○	○
I understand that information I provide will be used for a Bachelor's Thesis and possibly for a publication. (4)	○	○
I understand that personal information collected about me that can identify me, such as my name or where I live, will not be shared beyond the study team. (5)	○	○
I agree to be audio and video recorded. (6)	○	○
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. (7)	○	○

Q7 I hereby declare that I am at least 18 years old, have read the above information and that I voluntarily participate in this study.

○ Yes, I hereby declare that I am at least 18 years old, have read the above information and agree to participate voluntarily. (1)

○ No, I would like to end this session (2)

Appendix B: Questionnaires

Appendix B1: Initial Trust questionnaire

Disposition to Trust (McKnight et al., 2002) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

1 In general, people do care about the well-being of others.

2 The typical person is sincerely concerned about the problems of others.

3 Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.

4 In general, most folks keep their promises.

5 I think people generally try to back up their words with their actions.

6 Most people are honest in their dealings with others.

7 I usually trust people until they give me a reason not to trust them.

8 I generally give people the benefit of the doubt when I first meet them.

9 My typical approach is to trust new acquaintances until they prove I should not trust them.

Personal Innovativeness (adapted from McKnight et al., 2002) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

1 I like to explore new Web sites/technologies.

2 When I hear about a new Web site/technology, I often find an excuse to go visit/use it.

3 Among my peers, I am usually the first to try out new Internet sites/technologies.

4 In general, I am not interested in trying out new Web sites/technologies.

5 When I have some free time, I often explore new Web sites/technologies.

Appendix B2: 15-item Chatbot Usability Scale (CUS)

15-item Chatbot Usability Scale (CUS) (Borci et al, Under Review) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

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

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 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 chatbot's 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 B3: UMUX-LITE

UMUX-LITE (Lewis et al., 2013) (5-pt Likert scale – 1-strongly disagree, 5-strongly agree)
Answer the following questions based on your experience with the chatbot:

1	The chatbot's capabilities meet my requirements.
2	The chatbot is easy to use.

Appendix B4: Trust questionnaire after the chatbot interaction

Technology trusting performance (Lankton et al., 2014) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

Based on your experience with the chatbot it:

1 had the functionality I needed.

2 had the features required for my tasks.

3 had the overall capabilities I needed.

4 provided the help I needed to complete tasks successfully.

5 provided competent guidance through a help function.

6 supplied my need for help through a help function.

7 did not fail me.

8 did not malfunction for me.

9 provided error-free results.

Technology trusting performance (adapted from Lankton et al., 2014) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

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

1 When I have a tough task, I feel I can depend on the chatbot.

2 I can always rely on the chatbot in completing a tough task.

3 The chatbot is a product on which I feel I can fully rely when working on an essential task.

4 I feel I can count on the chatbot when working on an important task.

Usefulness (Lankton et al., 2014) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

Based on your experience with the chatbot it:

1 increased my productivity.

2 improved my performance.

3 enhanced my effectiveness.

4 was useful.

Appendix B5: Usage Continuance Intention questionnaire

Usage continuance intention (adapted from Lankton et al., 2014) (5-pt Likert scale – 1-totally disagree, 5-totally agree)

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

1 In the near future, I intend to continue using the chatbot.

2 I intend to continue using the chatbot to find intended information.

3 I plan to continue using the chatbot.

Appendix C: List of chatbots and tasks

1. https://manychat.com/blog/

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.

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.

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.

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.

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.

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.

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.

8. https://www.benefitcosmetics.com/en-us
Perform the following task using the chatbot:
You are interested in buying a brown mascara. Find out what options there are.

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.

10. https://www.absolut.com/en/
Perform the following task using the chatbot:
You are interested in finding out where the Absolut is from.

Appendix D: R script

```
################################
### Bachelor Thesis analysis ###
################################

###packages to install- only needs to be done once.

install.packages("readxl")      ### reads excel
install.packages("lavaan")      ### does LAtent VAriable ANalysis see
https://lavaan.ugent.be/
install.packages("lavaanPlot")  ### make plots https://cran.r-
project.org/web/packages/lavaanPlot/vignettes/Intro_to_lavaanPlot.html
install.packages("dplyr")
install.packages("haven")
install.packages("ggpubr")
install.packages("semPlot")

###pull packages out of the library

library(readxl)
library(lavaan)
library(dplyr)
library(haven)
library(ggpubr)
library(knitr)
library(semPlot)
library(psych)

###turn off scientific notation

options(scipen = 999)

###read in the data set
```

```
BUS_CFA <- read_excel("Working Data set2.xlsx")
View(BUS_CFA)



############################# RQ1
##############################################

###########
### CFA ###
###########



#F1 - Perceived accessibility to chatbot functions
#f2 - Perceived quality of chatbot functions
#f3 - Perceived quality of conversation and information provided
#f4 - Perceived privacy and security
#f5 - Time response

#Normality check: RESULTS NOT NORMALLY DISTRIBUTED
shapiro.test(BUS_CFA$USIC_1)
shapiro.test(BUS_CFA$USIC_2)
shapiro.test(BUS_CFA$USIC_3)
shapiro.test(BUS_CFA$USIC_4)
shapiro.test(BUS_CFA$USIC_5)
shapiro.test(BUS_CFA$USIC_6)
shapiro.test(BUS_CFA$USIC_7)
shapiro.test(BUS_CFA$USIC_8)
shapiro.test(BUS_CFA$USIC_9)
shapiro.test(BUS_CFA$USIC_10)
shapiro.test(BUS_CFA$USIC_11)
shapiro.test(BUS_CFA$USIC_12)
shapiro.test(BUS_CFA$USIC_13)
shapiro.test(BUS_CFA$USIC_14)
shapiro.test(BUS_CFA$USIC_15)


### Model1:Five Factors: There is a warning variance is negative it is a sign that the model is not good enough and there is too much collinearity (Heywood Case). Variance table suggest to remove USIC_1

model1 <- 'f1 =~ USIC_1 + USIC_2
        f2 =~ USIC_3 + USIC_4 + USIC_5 + USIC_6 + USIC_7 + USIC_8 + USIC_9
        f3 =~ USIC_10 + USIC_11 + USIC_12 + USIC_13
        f4 =~ USIC_14
        f5 =~ USIC_15'

fit <- cfa(model1, data = BUS_CFA, estimator="MLR", mimic="Mplus")
summary(fit, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE)



### Model M2:Removing Item 1 (the model could be improved looking at the residuals an modification indexes)

model2<-'f1 =~ USIC_2
        f2 =~ USIC_3 + USIC_4 + USIC_5 + USIC_6 + USIC_7 + USIC_8 + USIC_9
        f3 =~ USIC_10 + USIC_11 + USIC_12 + USIC_13
        f4 =~ USIC_14
        f5 =~ USIC_15'

fit2 <- cfa(model2, data = BUS_CFA, estimator="MLR", mimic="Mplus")
summary(fit2, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE) #using the proposed cutoff criteria, the ML-based TLI, Mc, and RMSEA tend to overreject true-population models at small sample size and thus are less preferable when sample size is small
```

```
# Residual
resid(fit2, type = "normalized")$cov %>% kable(caption = "Normalized Residual Variance-
Covariance Matrix",digits = 3)

# Modification index
modindices(fit2, minimum.value = 10, sort = TRUE)%>% slice(1:30) %>% kable(caption =
"Modification Indices", digits = 3)



### Model3: New factor items USIC_11 and USIC_12

model3<-'f1 =~ USIC_2
        f2 =~ USIC_3 + USIC_4 + USIC_5 + USIC_6 + USIC_7 + USIC_8 + USIC_9
        f3 =~ USIC_10 + USIC_13
        f3a =~ USIC_11 + USIC_12
        f4 =~ USIC_14
        f5 =~ USIC_15'

fit3 <- cfa(model3, data = BUS_CFA, estimator="MLR", mimic="Mplus")
summary(fit3, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE) #using the
proposed cutoff criteria, the ML-based TLI, Mc, and RMSEA tend to overreject true-
population models at small sample size and thus are less preferable when sample size is small


# Residual
resid(fit3, type = "normalized")$cov %>% kable(caption = "Normalized Residual Variance-
Covariance Matrix",digits = 3)

# Modification index
modindices(fit3, minimum.value = 10, sort = TRUE)%>% slice(1:20) %>% kable(caption =
"Modification Indices", digits = 3)
```

```
### Model4: New factor items USIC_11 and USIC_12; drop USIC_4

model4<-'f1 =~ USIC_2
        f2 =~ USIC_3 + USIC_5 + USIC_6 + USIC_7 + USIC_8 + USIC_9
        f3 =~ USIC_10 + USIC_13
        f3a =~ USIC_11 + USIC_12
        f4 =~ USIC_14
        f5 =~ USIC_15'

fit4 <- cfa(model4, data = BUS_CFA, estimator="MLR", mimic="Mplus")
summary(fit4, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE) #using the proposed cutoff criteria, the ML-based TLI, Mc, and RMSEA tend to overreject true-population models at small sample size and thus are less preferable when sample size is small


# Residual
resid(fit4, type = "normalized")$cov %>% kable(caption = "Normalized Residual Variance-Covariance Matrix",digits = 3)

# Modification index
modindices(fit4, minimum.value = 10, sort = TRUE)%>% slice(1:20) %>% kable(caption = "Modification Indices", digits = 3)



### Model5: New factor items USIC_11 and USIC_12; Covariation USIC_3~~USIC_9, USIC_5~~USIC_6, USIC_6~~USIC_8

model5<-'f1 =~ USIC_2
        f2 =~ USIC_3 + USIC_5 + USIC_6 + USIC_7 + USIC_8 + USIC_9
        f3 =~ USIC_10 + USIC_13
        f3a =~ USIC_11 + USIC_12
        f4 =~ USIC_14
        f5 =~ USIC_15
        USIC_3~~USIC_9
        USIC_5~~USIC_6
        USIC_6~~USIC_8'

fit5 <- cfa(model5, data = BUS_CFA, estimator="MLR", mimic="Mplus")
summary(fit5, standardized=TRUE, ci=TRUE, fit.measures=TRUE, rsq=TRUE) #using the
proposed cutoff criteria, the ML-based TLI, Mc, and RMSEA tend to overreject true-
population models at small sample size and thus are less preferable when sample size is small


# Residual
resid(fit5, type = "normalized")$cov %>% kable(caption = "Normalized Residual Variance-
Covariance Matrix",digits = 3)

# Modification index
modindices(fit5, minimum.value = 10, sort = TRUE)%>% slice(1:10) %>% kable(caption =
"Modification Indices", digits = 3)


### Design structure of the model using SEM package
semPaths(fit5, "std")


#### Cronbach's alpha
install.packages("psych")
library(psych)

alpha(subset(BUS_CFA, select = c(USIC_2:USIC_3, USIC_5:USIC_15)), check.keys = T)
#0.92
```

```
################################ RQ 2
##########################################

###########################
### Correlation testing ###
###########################



###Spearman's rank-order correlation
cor.test(BUS_CFA$USIC.mean, BUS_CFA$UMUX.mean,
        method = "spearman")
cor.test(BUS_CFA$f1.mean, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$f2.mean, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$`f3.mean`, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$USIC_14, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$USIC_15, BUS_CFA$UMUX.mean, method = "spearman")

cor.test(BUS_CFA$USIC.II.mean, BUS_CFA$UMUX.mean,
        method = "spearman")
cor.test(BUS_CFA$USIC_2, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$f22.mean, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$f33.mean, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$f3a.mean, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$USIC_14, BUS_CFA$UMUX.mean, method = "spearman")
cor.test(BUS_CFA$USIC_15, BUS_CFA$UMUX.mean, method = "spearman")

#visual representation
plot(BUS_CFA$USIC.mean, BUS_CFA$UMUX.mean, xlab = "Meanscore 15-items CUS",
ylab = "Meanscore UMUX-LITE")
plot(BUS_CFA$USIC.II.mean, BUS_CFA$UMUX.mean, xlab = "Meanscore 14-items
CUS", ylab = "Meanscore UMUX-LITE")
```

```
############################ RQ 3-5
############################################

############################
#### Mixed effect models ###    #good for higher sample size,
############################    #account for correlation between data coming from 3rd
party [here: Chatbots],
                                #estimate fewer parameters
                                #& avoid problems with multiple comparisons

### install packages
install.packages("ggplot2")
install.packages("lme4")
install.packages("nlme")
install.packages("MuMIn")

### call packages
library(ggplot2)
library(lme4)
library(nlme)
library(MuMIn)


### Standardize data --> mean = 0 & sd = 1 for all
DispTrust2 <- scale(BUS_CFA$DispTrust_Mean)
PersInno2 <- scale(BUS_CFA$PersInno_Mean)
Sat2 <- scale(BUS_CFA$Sat_Mean)
TruAft2 <- scale(BUS_CFA$TruAft_Mean)
Conti2 <- scale(BUS_CFA$Conti_Mean)


### fit model   -->  the former variable = y; the latter = x
Sat_DispTrust.lme <- lme(Sat2 ~ DispTrust2, random =~ 1|Chatbot, data = BUS_CFA)
summary(Sat_DispTrust.lme)       # NOT sign
```

```
TruAft_DispTrust.lme <- lme(TruAft2 ~ DispTrust2, random =~ 1|Chatbot, data = BUS_CFA)
summary(TruAft_DispTrust.lme)    # NOT sign

Sat_PersInno.lme <- lme(Sat2 ~ PersInno2, random =~ 1|Chatbot, data = BUS_CFA)
summary(Sat_PersInno.lme)        # sign, p = 0.0118 (< 0.05), t = 2.530, df = 389, std.err = 0.046, value = 0.116
r.squaredGLMM(Sat_PersInno.lme)  #HOWEVER: value very small (0.121) --> not much impact

TruAft_PersInno.lme <- lme(TruAft2 ~ PersInno2, random =~ 1|Chatbot, data = BUS_CFA)
summary(TruAft_PersInno.lme)     # NOT sign

TruAft_Sat.lme <- lme(TruAft2 ~ Sat2, random =~ 1|Chatbot, data = BUS_CFA)
summary(TruAft_Sat.lme)          # VERY sign - < 0.001, t = 38.717, df = 389, std.err = 0.023, value = 0.892
r.squaredGLMM(TruAft_Sat.lme)

Sat_TruAft.lme <- lme(Sat2 ~ TruAft2, random =~ 1|Chatbot, data = BUS_CFA)
summary(Sat_TruAft.lme)          # VERY sign - < 0.001, t = 38.473, df = 389, std.err = 0.023, value = 0.884
r.squaredGLMM(Sat_TruAft.lme)

Conti_TruAft.lme <- lme(Conti2 ~ TruAft2, random =~ 1|Chatbot, data = BUS_CFA)
summary(Conti_TruAft.lme)        # VERY sign - < 0.001, t = 24.867, df = 389, std.err = 0.031, value = 0.780
r.squaredGLMM(Conti_TruAft.lme)

Conti_Sat.lme <- lme(Conti2 ~ Sat2, random =~ 1|Chatbot, data = BUS_CFA)
summary(Conti_Sat.lme)           # VERY sign - < 0.001, t = 20.859, df = 389, std.err = 0.035, value = 0.723
r.squaredGLMM(Conti_Sat.lme)
```

```
### Check assumptions

# fit ~ resid  --> equally beneath and above line
plot(Sat_DispTrust.lme)      # ok
plot(Sat_PersInno.lme)       # ok
plot(TruAft_DispTrust.lme)   # ok
plot(TruAft_PersInno.lme)    # ok
plot(TruAft_Sat.lme)         # ok
plot(Sat_TruAft.lme)         # ok
plot(Conti_TruAft.lme)       ### neg diagonal
plot(Conti_Sat.lme)          ### neg diagonal

# qqplot  --> should be in line
qqnorm(resid(Sat_DispTrust.lme))
qqline(resid(Sat_DispTrust.lme))    # ok

qqnorm(resid(Sat_PersInno.lme))
qqline(resid(Sat_PersInno.lme))     # ok

qqnorm(resid(TruAft_DispTrust.lme))
qqline(resid(TruAft_DispTrust.lme)) # ok

qqnorm(resid(TruAft_PersInno.lme))
qqline(resid(TruAft_PersInno.lme))  # ok

qqnorm(resid(TruAft_Sat.lme))
qqline(resid(TruAft_Sat.lme))       # ok

qqnorm(resid(Sat_TruAft.lme))
qqline(resid(Sat_TruAft.lme))       # ok

qqnorm(resid(Conti_TruAft.lme))
qqline(resid(Conti_TruAft.lme))     # ok
```

```
qqnorm(resid(Conti_Sat.lme))
qqline(resid(Conti_Sat.lme))        # ok

### Visual representation of sign relations
plot(BUS_CFA$PersInno_Mean, BUS_CFA$Sat_Mean, xlab = "Personal Innovation", ylab = "Satisfaction")

plot(BUS_CFA$Sat_Mean, BUS_CFA$TruAft_Mean, xlab = "Satisfaction", ylab = "Trust after interaction")
plot(BUS_CFA$TruAft_Mean, BUS_CFA$Sat_Mean, xlab = "Trust after the interaction", ylab = "Satisfaction")

plot(BUS_CFA$Conti_Mean, BUS_CFA$TruAft_Mean, xlab = "Usage continuance intention", ylab = "Trust after interaction")
plot(BUS_CFA$Conti_Mean, BUS_CFA$Sat_Mean, xlab = "Usage continuance intention", ylab = "Satisfaction")
```

Appendix D1: Confirmatory factor analyses results

Table D1.1

Normalized Residual Variance-Covariance Matrix of Model M2 for CUS_10 to CUS_13

	CUS_10	**CUS_11**	**CUS_12**	**CUS_13**
CUS_10	0.000	-0.080	0.319	0.112
CUS_11	-0.080	0.000	1.719	-0.416
CUS_12	0.319	1.719	0.000	-0.145
CUS_13	0.112	-0.416	-0.145	0.000

Table D1.2

Normalized Residual Variance-Covariance Matrix of Model M3 for CUS_3 to CUS_9

	CUS_3	**CUS_4**	**CUS_5**	**CUS_6**	**CUS_7**	**CUS_8**	**CUS_9**
CUS_3	0.000	-0.385	-0.017	-0.136	-0.185	-0.628	0.899
CUS_4	-0.385	0.000	0.190	-0.563	0.482	-0.526	0.658
CUS_5	-0.017	0.190	0.000	1.505	-1.102	1.869	-0.617
CUS_6	-0.136	-0.563	1.505	0.000	-0.979	1.574	-1.323
CUS_7	-0.185	0.482	-1.102	-0.979	0.000	0.145	1.366
CUS_8	-0.628	-0.526	1.869	1.574	0.145	0.000	-1.520
CUS_9	0.899	0.658	-0.617	-1.323	1.366	-1.520	0.000

Table D1.3

Normalized Residual Variance-Covariance Matrix of Model M4 for CUS_3 to CUS_9

	CUS_3	**CUS_5**	**CUS_6**	**CUS_7**	**CUS_8**	**CUS_9**
CUS_3	0.000	-0.079	-0.263	-0.146	-0.731	0.920
CUS_5	-0.079	0.000	1.425	-1.053	1.809	-0.581
CUS_6	-0.263	1.425	0.000	-0.953	1.457	-1.313
CUS_7	-0.146	-1.053	-0.953	0.000	0.173	1.446
CUS_8	-0.731	1.809	1.457	0.173	0.000	-1.505
CUS_9	0.920	-0.581	-1.313	1.446	-1.505	0.000

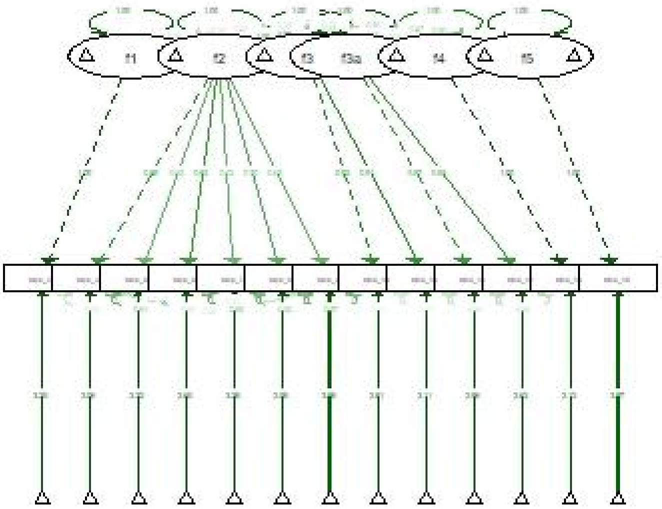

Figure D1 Visual representation of the modified six-factor model M5 of the confirmatory factor analysis

Appendix D2: Visual representations of the significant regression correlations

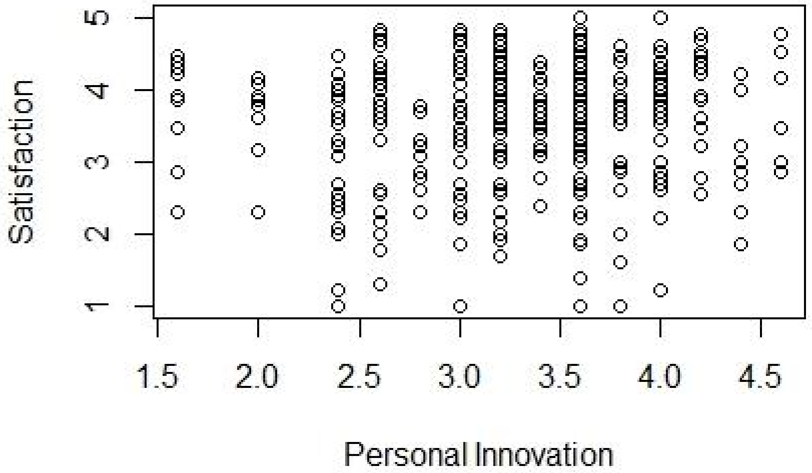

Figure D2.1 Visual representation of effect of Personal innovativeness on Satisfaction

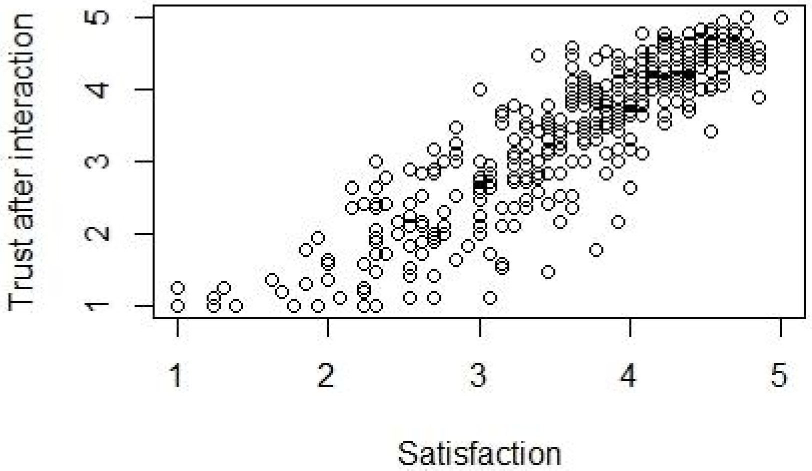

Figure D2.1 Visual representation of the effect of Satisfaction on Trust after the interaction

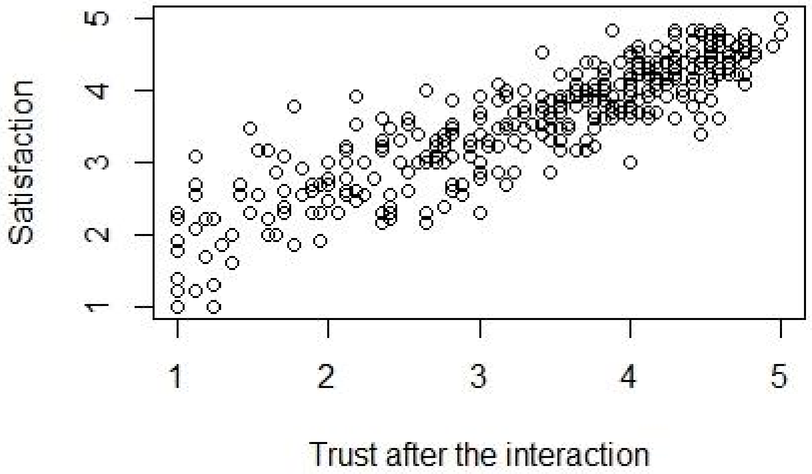

Figure D2.3 Visual representation of the relationship between Trust after the interaction with satisfaction

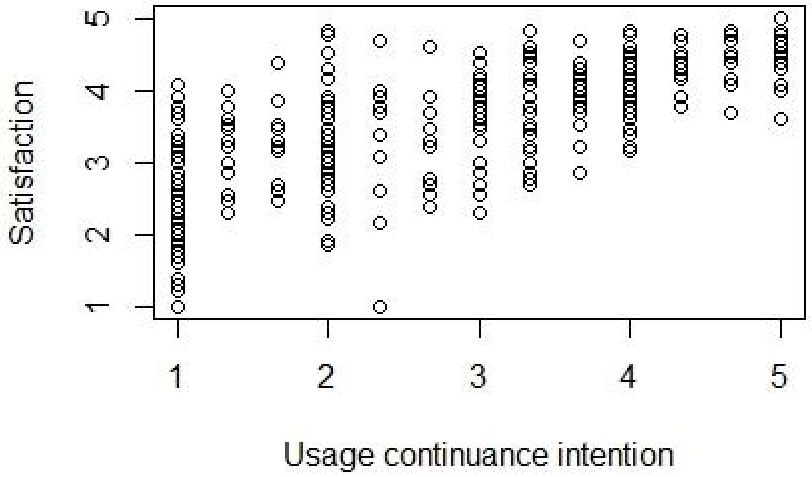

Figure D2.4 Visual representation of the relationship Usage continuance intention and satisfaction

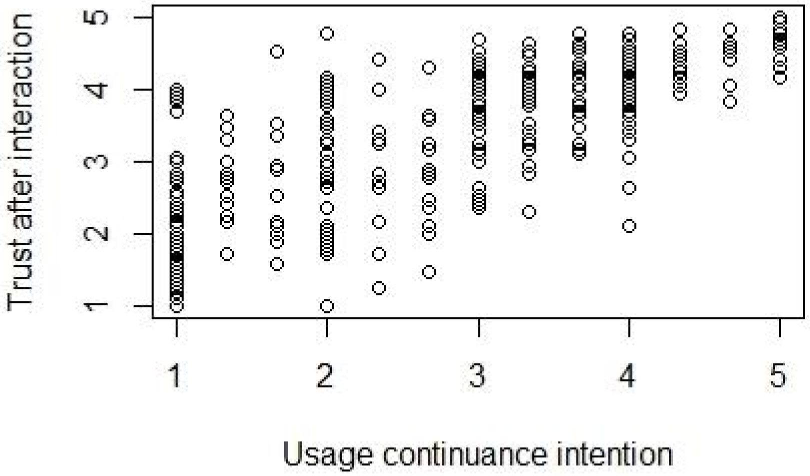

Figure D2.5 Visual representation of the effect of Usage continuance intention on Trust after the interaction

Appendix E: Modified 13-items CUS model

Table E.1

Modified Chatbot Usability Scale

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