BACHELOR THESIS

Confirmatory Factor Analysis of a new Satisfaction Scale for conversational agents and the role of decision-making styles

Sabin Manuela Kerwien Lopez

2154196

First Supervisor: dr. S. Borsci

Second Supervisor: prof. dr. F. van der Velde

University of Twente

Faculty of Behavioural, Management and Social Sciences

Department of Cognitive Psychology and Ergonomics

2021

Abstract

Companies working in consumer service are increasingly implementing the usage of chatbots on their websites, to help the users to reach their end goal. In many instances, however, the interactions do not meet the expectations the users have towards chatbots. It is therefore important to have a measure of user satisfaction to assess whether the chatbot can be improved or is able to fulfil its task according to the expectations. A standardized measure for satisfaction levels in chatbot interaction is not yet readily available. Borsci et al. (under review), took the first step to develop such a questionnaire. More research is, however, needed to assess its psychometric properties and its correlation to other standardized measures of satisfaction. Participants were invited to interact with 10 different chatbots and their satisfaction levels were measured after each interaction. Thereafter, confirmatory factor analysis was performed and results show evidence for a new four-factor model, consisting of fourteen items. This new model is reliable, as further reliability analysis showed. Correlation analysis with the UMUX-Lite questionnaire showed high and significant correlation results, indicating a good external validity. Moreover, to enable new populations to use this questionnaire, the scale was translated into Spanish and correlation analysis with the English version indicated that the translation was reliable and measures a similar concept of satisfaction. Lastly, the overall influence of decision-making styles, as measured by the general decision-making style scale, on satisfaction levels was assessed. Results showed that decision-making styles did not significantly influence satisfaction levels measured by the new scale.

Keywords: Artificial Intelligence, Chatbots, Conversational Agents, BotScale, Satisfaction, Decision-Making Style

Table of contents

1. Introduction

1.1 Conversational Agents

Exactly 71 years ago, in 1950, Alan Turing was already speculating on the future of computers but more specifically asking whether computers will be able to communicate similarly to human beings. He concluded with the idea that in the near future this would be possible (Zemčík, 2019). A specific program that focuses on this question of communication with humans are chatbots, also known as conversational agents. The term "chatbot" consists of the words "chat" and "robot", which essentially entails the definition of such systems. A chatbot is, hence, defined as an artificial intelligence software that performs a conversation by, more specifically, simulating human language (Sanny et al., 2020). Fundamentally, it is a computer program that uses text-based language as input while successively creating natural language output (Valerio et al., 2017). Due to their nature, they enable humans to interact with them (Valerio et al., 2017). Although the more frequent application of such chatbots can be seen in recent years, chatbots were already developed in the 1960s (Khan, 2018). The ongoing process of development, especially regarding natural language interpretation, resulted in various chatbot software, some of which employ simple abstractions and others that employ more complex concepts (Paikari & Van der Hoek, 2018). Thus, there are two different types of chatbots, the main distinction is made between rule-driven conversational agents and chatbots that are based on artificial intelligence.

The first type of chatbots are keyword recognition-based and are, therefore, monitoring user input. In that sense, they are listening to what the user is saying. Thereby, they search for and recognize patterns to then deliver pre-defined answers to those questions (Bieliauskas & Schreiber, 2017; Io & Lee, 2017). Due to this pre-defined nature, open conversations are not possible. One specific problem area for this type of chatbot is when users use sentences that entail redundant keywords, as these will trigger unneeded and false responses (Gupta et al., 2020). The second type of AI-based chatbots is also known as contextual chatbots (Gupta et al., 2020). These complex versions of chatbots aim at enabling engagement that is human-like and intelligent. Furthermore, these chatbots aim at interpreting the user's goal and meaning

within this interaction and thereafter to give the needed information to reach this goal. In comparison to the previously described type, these chatbots go further by learning from experience with each interaction done (Io & Lee, 2017). Thereby, with each interaction, they improve both their understanding of user input as well as the accuracy of their responses. Algorithms are used to create a meaningful output of the data that is gathered in each conversation by, for example, connecting ideas and themes (Soni, 2018). This can be done by using a mixture of machine learning and AI to understand the needs of the customer (Soni, 2018).

Companies are increasing the use of such conversational agents to supply information to the user (Khan, 2018). They are predominantly used in the domain of customer service and experience (Sanny et al., 2020). Analyzing the reasons for the increase in their popularity, two main reasons can be found. Valério et al. (2017) suggest that the advancing developments in the ease of implementation account for this popularity. As of 2016 for example Facebook introduced their messenger application programming interface which allows for the simple and fast creation of personalized chatbots (Khan, 2018). Their primary usage in customer service accounts for a further aspect of their popularity, which is based on their ability to add a personal channel of communication and to provide real-time service (Adam et al., 2020; Følstad & Brandtzaeg, 2020). As Xu et al., (2017) suggest, the usage of chatbots offers the possibility of replacing or altering customer service. They enable 24-hour support and more importantly offer this support regardless of the customers geographic location (Ashfaq et al., 2020). An additional factor contributing to this is their ability to converse in a human-like manner with consumers (Pfeuffer et al., 2019). Customers are hence able to receive unrestricted support that simultaneously offers personalized conversations (Zumstein & Hundertmark, 2018). Simulating human conversations, however, is not the end goal of the implementation of chatbots. Conversational agents are implemented to enable users to achieve a certain goal and to receive the information that is needed to reach this objective (Følstad & Brandtzaeg, 2020). This can vary from getting information about certain products to placing orders for products or booking activities (Ashfaq et al., 2020). Their successful use, however, demands the correct

implementation and the consequent satisfaction on behalf of the user. Therefore, research in the area of interaction with chatbots is needed.

1.2 Interaction with chatbots

Human-Computer Interaction (HCI) is a research discipline that studies the way humans interact with computers and other technologies (Oulasvirta & Hornbæk, 2016; Bevan, 2001). Research in this area explores for example the motivation of people to use chatbots (Brandtzaeg & Følstad, 2017). Other research focuses on differences in the conversations between humans and conversations between chatbots and humans (Hill & Farreras, 2015). Results showed that human individuals tend to imitate human-human conversations in their interaction with chatbots, with some difference in the length of the conversation due to the technological nature of chatbots. The two areas of extensive research are the areas of usability and user experience with chatbots (Holtgravers et al., 2007; Arujo, 2018; Gnewuch et al., 2017).

An essential concept of HCI is thus usability (Bevan, 2001), which is defined as the extent to which a user can use a certain product to achieve his goal effectively, efficiently and in a satisfactory way in a specific context of use (ISO 9241-11, 2018). This concept of usability and its three metrics can be further transferred to usability testing. The main aim is concerned with enabling a researcher to assess a certain product on the basis of the aforementioned metrics. In that sense, these three metrics can be used to measure the usability of a certain product (Ferreira et al., 2020). The gathered information can be used to see in what way the product can be enhanced in terms of user usability. This process requires the researcher to develop tasks that the user has to complete and consequently measure the metrics of effectiveness, efficiency, and satisfaction. Joo (2010) proposed the idea that these three metrics are highly correlated with each other. The degree of correlation, however, depends on influencing variables such as the context of use, task complexity, measures used or the domain that is the topic of research (Frøkjær, Hertzum, & Hornbæk, 2000; Hornbæk & Law, 2007). One distinction for example, as proposed by Frøkjær et al. (2000), can be made when dealing with routine tasks, results on efficiency and effectiveness

tend to be higher than on novel tasks. This is explained by the automation and practice of such actions. When dealing with user experiences or when wanting to assess subjective measures, the variable of satisfaction is the most crucial measure in research (Hassan & Galal-Edeen, 2017).

To comprehend the measures needed for effective usability testing as well as the differences between the variables the definitions of the three metrics are presented. Effectiveness relates to the extent to which a user is able to accurately and completely accomplish a goal (ISO 9241-11, 2018). Efficiency deals with the resources used when completing a certain task, thereby it analyzes the time invested in accomplishing a task (ISO 9241-11, 2018). Lastly, satisfaction is defined as "the extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (ISO 9241-11, 2018). In that sense, it also entails the comfort as well as the positive attitude a user has towards a system (Frøkjær et al., 2000). Its nature makes it inherently difficult to quantify as it is subjective, yet it is frequently applied to determine the success of a certain product (Feine, Morana & Gnewuch, 2019). Hence, it is a crucial factor in usability testing as it focuses on the subjective user's perception (Bevan, 2009). In this regard it focuses on the user experience, a subcategory of usability, thereby completing the assessment of the usability of a product (Hassan & Galal-Edeen, 2017). This principle of satisfaction will be the main focus of this study.

1.3 The necessity of a metric to assess the satisfaction of end-users during the interaction with chatbots

Luger and Sellen (2016) argue that the implementation of chatbots is often not in accordance with the expectations of the users. More specifically, users often report unsatisfactory interactions with chatbots. These include meaningless and illogical responses and therefore no usability of the information (Brandtzaeg & Følstad, 2017). Other users report a lack of empathy or sensitivity towards the user (Ashfaq et al., 2020). Such negative encounters might hinder the further development of chatbots and their implementation, regardless of their advantages, as users are less inclined to use them (Adam et al., 2020).

As Følstad and Brandtzaeg (2020) add, user experience needs to be improved to enable positive encounters and therefore increase the probability of users turning to chatbots for help. Su (1992) adds to this idea by suggesting that when dealing with information retrieval systems, such as conversational agents, satisfaction can be seen as an approach to measure the performance of and user experience with such systems. The primary usage of chatbots in consumer experience lays the foundation for the need to assess the extent to which users are satisfied with the specific chatbot. This is because conversational agents are often seen as dynamical additions to the experience of the consumer on the website of brands. As chatbots interact directly with potential consumers, their performance needs to satisfy the consumer, as high satisfaction values are highly and positively correlated to the success of a company (Oliver, 2010). As Feine, Morana and Gnewuch (2019) propose, in the context of consumer experience fast assessment of the satisfaction of the user is needed to ensure that customers don't have negative experiences. Thong and Yap (1996) lay the first idea of how to guarantee high satisfaction levels. They suggest that if a system meets the requirement a user has toward such a system the level of satisfaction will increase (Thong & Yap, 1996). Lewis (1995) adds to this idea and suggests that customers want usable products. Such outcomes can be enhanced by researching variables influencing user satisfaction, such as decision-making styles for example, and the results can thereafter be used to tailor the chatbots according to the user 's expectations and preferences (Kazeminia et al., 2019).

1.3.1 A satisfaction scale for chatbots

The main problem researchers are confronted with when wanting to measure satisfaction in the interaction with chatbots is that there is no readily available quantitative measure developed for the interaction with chatbots.

As researchers are confronted with this challenge new approaches to measure satisfaction were developed. The most popular being standardized scales of satisfaction to enable quantitative measurements (Kondo, 2001). A closer look at existing studies measuring satisfaction in the interaction with chatbots reveals that this leads to different measurements being used. In that sense, some studies

employ and modify existing scales that measure customer satisfaction (Chung et al., 2018; Eren, 2021). One specific example is the System Usability Scale (SUS), which consists of 10 items. As Sauro and Lewis' (2009) meta-review on post-hoc satisfaction questionnaires showed, this specific type of satisfaction scale was used in 43% of the studies, illustrating its popularity in academic research. A further popular example is the Usability Metric for User Experience (UMUX) which consists of four items the user has to answer. A newer ultrashort scale is available, the UMUX-Lite, which consists of two items. All three have excellent psychometric properties in the sense that they are both reliable and valid and correlate with each other (Lewis, Utesch & Maher, 2013; Borsci et al. 2015). The UMUX-Lite scale specifically has high-reliability values of $\propto$=.82 (Lewis, Utesch & Maher, 2013). The study by Lewis et al. (2013) showed further that it resulted in similar results as and correlated highly with the established and standardized SUS questionnaire (r=.81). At the same time, the UMUX Lite offers the advantage of being shorter and thus less restraining on the user. Longer questionnaires run the risk of participants experiencing response fatigue and thus bias the results (Helton, 2004). This risk runs especially when studies are long or use repeated measures.

Such standardized measurements can be used but hold the disadvantages of not being developed specifically for the interaction with chatbots. One specific risk that arises then is the risk that certain valuable factors needed in the assessment of interaction with chatbots are not included in such general standardized measures (Tarverdiyeva & Borsci, 2019). Other researchers measured satisfaction based on various factors, such as perceived empathy or helpfulness (Xu et al., 2017; Heller et al., 2005). Maroengsit et al., (2019) proposed yet a different approach of evaluating satisfaction in two levels, the first evaluation is done on the whole conversation while the second is done for each interaction individually. Concludingly it can be said that there are various measures of satisfaction but that none of these was developed specifically for the interaction between humans and chatbots. This is especially problematic as satisfaction levels with chatbot interactions are not measured in a similar way. As Baroudi and Orlikowski (1988) argue standardized measures offer the advantage of being widely applicable and they reduce time

investment as a readily available measure can be used. Thereby the authors are stressing the importance of a standardized scale for measuring satisfaction levels in interactions with chatbots.

To counteract this problem of measuring satisfaction within the specific context of interaction with chatbots, a new satisfaction scale, the BotScale (Borsci et al. , 2021, under review), was developed. The first version of the BotScale was developed using a systematic literature review and it consisted of 42 items (Tarverdiyeva & Borsci, 2019). A second literature review was conducted alongside the consultancy of experts with chatbots to review the existing scale and its factors (Balaji & Borsci, 2019) Thereafter a focus group was included in the research and asked to evaluate the items and factors that were deemed important for the newly developed scale (Balaji & Borsci, 2019). Balaji and Borsci concluded their research by assessing the scale by letting participants interact with chatbots and thereafter asking them to fill in the scale. Results indicated that a shortened version of 14-items and four factors showed improved results over the 42-item questionnaire. A second study was used to replicate this model and results showed evidence for the four-factor model (Silderhuis & Borsci, 2020). A more extensive review conducted by Borsci et al., (2021, under review) showed that the original scale could be reduced to a final set of 15 items by exploratory factorial analysis, divided into 5 underlying factors, this questionnaire is also known as the BotScale (see also: Appendix A). This five-factor model showed high reliability ($\propto$=0.87). Nevertheless, a confirmatory factorial analysis was not performed on the final version. A confirmatory factor analysis, however, is crucial to test whether the items of a questionnaire correctly measure the hypothesized factor structure (Holye, 2000). Therefore, further research on the BotScale psychometric properties is needed to assess whether, the final version, is a reliable and valid measure of satisfaction.

Moreover, the BotScale was initially developed in English and subsequently translated and validated into a Dutch version (van den Bos & Borsci, 2021). Translating a survey questionnaire into new languages offers the main advantage of offering the possibility of access to a larger population, as users are able to complete the questionnaire in their own language (Presser et al., 2004). In that sense, one can assume that the scale will yield more reliable results if it is completed in the mother language of the participant (Banville, Desrosiers, & Genet-Volet, 2000).

1.4 The Relationship between satisfaction and Decision-Making Styles

The relation between decision making, the use of chatbots as well as user satisfaction has been a topic of research. Decision-making styles are generally defined as common patterns that humans take to come to such decisions (Raffaldi et al., 2012). A study by Alavi et al. (2016) illustrated the importance of analysing decision-making styles in the context of consumer experience. Their results showed that decision-making styles can be used as a conjecture of satisfaction levels (Alavi et al., 2016). A further step can be taken by applying the relationship between satisfaction and decision-making styles to the context of chatbots. One reason for the emergence of this area of research is that chatbots have the fundamental task of helping in the decision-making process (De Vreede, Raghavan, & De Vreede, 2021). It, therefore, engages customers and offers organizational assistance as well as responses to specific questions. While determining the decision-making style of a person a further step can be taken to use this information to enhance interactions with chatbots. This works as one of the various functions of chatbots is to guide the user in their decision-making process (Shumanov & Johnson, 2020). Kazeminia et al., (2019) propose that a better comprehension of the relation between decision-making behaviour and satisfaction enables the personalization of chatbots (Kazeminia et al., 2019). This application can ultimately lead to chatbots that enable a positive experience by creating tailored chatbots, increasing consumer experience (Kaptein et al., 2010; Zhou et al., 2019; Bologna et al., 2013). Hence, one particular idea is that in order to increase levels of satisfaction the chatbot can be personalized according to the style of the user (Oliveira et al., 2013). This way tailored assistance to decision-making processes can be offered (Shumanov & Johnson, 2020). As Häubl and Trifts (2000) propose in the context of consumer service, decision support systems, that for example process or organize information, can be offered to the user to facilitate decision making according to personal preferences.

Previous work from Ciovati (2020), e.g, focused on maximizing theory as the underlying explanation of decision-making behaviour in individuals and the related level of satisfaction. The maximizing theory is used to study decision-making behaviour and proposed two types of behaviours: 1)

maximizers, who rationally make decisions and 2) satisfiers, who come to decisions based on their interests and intuitions. Results of this study showed that the two different decision-making styles yielded different levels of satisfaction. Maximizers or rational decision-makers tended to yield lower satisfaction levels as they continued searching for better alternatives (Ciovati, 2020). While maximizing theory can be used to assess decision-making style, there is a further scale that is extensive and superior to this theory, as it is encompassed and validated on various occasions (Fischer et al., 2015; Berisha, Pula, & Krasniqi, 2018).

Scott and Bruce (1995) developed this measurement of decision-making styles, which is known as the general decision-making style questionnaire (GDMS). The underlying reason for the development of the scale was that until then such a standardized and validated measure for measuring decision-making styles was not readily available (Scott & Bruce, 1995). The scale distinguishes between five dimensions of decision making: Rational, Avoidant, Dependent, Intuitive and Spontaneous Decision-Making Style. There is a general consensus of these five decision-making styles, with each individual having one particularly dominant style (Raffaldi et al., 2012). In that sense, it offers a broader spectrum of classification. The general decision-making style measures these five styles on a five-point Likert Scale (Loo, 2000). The scale has been successfully applied to various contexts, including an educational and military setting (Girard et al., 2016). The scale is especially successful as it has been validated in various countries and populations, among others Sweden, India, Canada and Spain (Thunholm, Verma & Rangnekar, 2015; 2004; Girard et al., 2016; Alacreu-Crespo et al., 2019). Its psychometric properties have been viewed as good (Kazeminia et al., 2019). Loo (2000) evaluated the psychometric properties of the scale and found moderate to good reliability indexes for each factor (Rational $\propto$=0.81, Intuitive $\propto$=0.79, Dependent $\propto$=0.62, Avoidant $\propto$=0.84, Spontaneous $\propto$=0.83). Researchers using the scale in the previously mentioned different populations have confirmed the underlying five-factor structure (Alacreu,-Crespo et al., 2019; Loo, 2000).

1.5 Aims of the present study

The present study aims at re-evaluating by a confirmatory factorial analysis (CFA) the psychometric properties of the BotScale. Additionally, this work aims to: i) propose a new translation of the scale in Spanish and ii) explore the influence of decision-making styles on the satisfaction of participants after interacting with chatbots measured by the BotScale. To achieve our goals first we will perform a confirmatory analysis of the BotScale to establish its factorial structure and its internal validity, as well as its external validity in terms of correlation with a classic satisfaction scale (UMUX Lite). Two research questions are associated with this goal:

RQ1: Can the factorial structure of the BotScale found in previous exploratory analyses be confirmed?"

RQ2: "Do the results from the Botscale correlate with the results of the UMUX Lite?"

To enlarge the usage possibilities of this questionnaire it is further necessary to translate and validate the questionnaire in additional languages. Before translating a questionnaire into a new language, two specific procedures are recommended. While the translation of questionnaires is often seen as an easy procedure Presser et al., (2004) propose that such translations are complex and time-consuming. The authors argue that for a thorough translation of a scale it is necessary to minimise the discrepancies between the versions. In the end, both questionnaires should ask the participant the same questions while communicating the same meaning. Procedures recommended are bilingual translators, and a team approach of at least two translators to enable a back translation into the original language (Presser et al., 2004). Therefore, we will translate the BotScale into Spanish, in accordance with the specific procedures proposed. This way other researchers have access to this metric and can conduct research in different languages with a validated and standardized measurement. To expand the potential use of this scale we will check the quality of the translation by assessing the psychometric properties of the translated scale, in line with this research question:

RQ3: "Does the Spanish translation of the BotScale present similar psychometric properties to the original version?"

Finally, this present study aims at researching if the decision-making style measured by GDMS affects satisfaction during the interaction with chatbots measured by the Bot Scale.

RQ4: "Does the decision-making style of an individual influence the level of satisfaction in users of chatbots?"

In line with previous studies on DM and user satisfaction, expectations are that decision-making styles will influence the satisfaction levels of users after the interaction with a chatbot. Previous studies focused solely on the rational decision-making style and found negative relationship with satisfaction levels. The negative relationship was explained because users tend to spend more time continuing to look for better options and are not easily satisfied with what is presented to them (Cheek & Schwartz, 2016). Thus, it is hypothesized that the results of this study will be similar, and a negative correlation between the rational decision-making style and satisfaction levels will be found.

2. Method

2.1 Design

The study employed a within-subject design with the independent variable of decision-making behaviour. The dependent variable in that sense was the satisfaction level of participants rated by the Bot Scale. Primary data was gathered through a survey. Participants were allowed to select their preferred language choosing between English, Spanish, and German. It also included German, as this work is part of a wider study to validate the BotScale in multiple languages. Thereafter an extensive confirmatory factor analysis was conducted to investigate the psychometric properties of the BotScale. Furthermore, correlation analyses were conducted to analyze the relationship between the translated versions,

2.2 Participants

Researchers used a convenience sample. Participants were primarily recruited within the circle of acquaintances of the researchers. Additionally, the study was also published on the website "Sona System" of the University of Twente, providing students with course credits for participation. In total, 74 entries were recorded in Qualtrics. Hereafter, all participants that did not complete the survey correctly were removed. This resulted in the exclusion of 19 entries and led to a total of 55 complete responses. The analysis consisted therefore of 55 participants. As each participant was asked to interact and assess ten chatbots, we collected a total of 550 BotScale questionnaires.

Thirty-four participants were female and twenty-one were male. The age ranged from 18 to 72 with m_{age}= 29.41 (SD = 13.99). All participants were fluent in English and additionally either in German or Spanish. Participants could freely choose the language they wanted to complete the survey in. The majority of participants were of German nationality, namely 38 participants, there were 2 Dutch participants and 15 selected "other" as their nationality. Nationalities included in this last category were Colombian, Greek, Salvadoran, American, Peruvian, Italian, Romanian, Vietnamese and Romanian.

2.3 Materials and Measures

Qualtrics. Qualtrics system was used to enable participants to interact with chatbots and answer the online questionnaire and thereby gather data from the participants.

Informed Consent. We requested the participants of the study to read and actively sign the informed consent (Appendix E). The informed consent contained information regarding the study, about the use of information gathered in the study, as well as contact information in case questions arose.

Demographic Questionnaire. Participants were asked for their (1) gender, (2) age, and (3) nationality. Starting from week three, we additionally asked them to fill in their full name and to complete a CAPTCHA test to ensure that robots did not make the responses.

General Decision-Making Style. The General Decision-Making Style Scale, developed by Scott and Bruce (1995) consists of 25 items which are measured on a five-point Likert Scale running from "strongly disagree" to "strongly agree" was used. The maximal scores indicated the dominant decision-making behaviour of an individual. Additionally, we used a validated Spanish version developed by Alacreu-Crespo et al., (2019; see Appendix D) for the participants taking the survey in the Spanish language.

UMUX Lite. After each interaction, two questionnaires were presented to the participants. First, we presented the UMUX Lite (Lewis et al., 2013), which consists of two items measured on a 7-point Likert Scale to the participants (Appendix C). Due to consistency, as all other scales employed a 5-point Likert scale, the researchers agreed to use this questionnaire with a 5-point Likert Scale.

Bot Scale. *S*atisfaction was further measured using the BotScale from Balaji and Borsci (2019) (Appendix A). This questionnaire comprises 15 items that are measured by a 5-point Likert Scale. This Scale was further translated into Spanish and used (Appendix B). The researcher completed the Spanish translation by first translating it from English to Spanish. An independent native Spanish speaker translated the scale back to English. There were no major differences in the back-translation.

Tasks. Participants had to complete one task per interaction with one conversation agent. In total, they had to interact with 10 web-based conversational agents. An overview of the tasks the participants

had to fulfil can be found in Appendix F. The sequence of chatbots was randomized. Four of these chatbots were already assessed by van den Bos and Borsci (2021). Consequently, this study introduced six new chatbots. The researcher provided participants with a link that redirects them to the webpage where the chatbots were implemented. Once the participant found the chatbot, they interacted with the chatbot and thereby completed the task.

2. 4 Procedure

Before starting and publishing the study ethical consent was requested from the BMS Ethics Committee of the University of Twente. The research was approved on the 7th of April. Thereafter the gathering of participants, which took place in two different ways, started. We generally invited participants to take part in the study by being contacted directly by the researcher or by selecting the link on the Sona System website.

During the first three weeks, researchers sent a link to the participants to join the online meeting. The participants then experimented in their private digital environment. Once the individual entered the session, we provided them with the link to the questionnaire on Qualtrics. Starting with week four participants were able to access the survey without supervision by the researchers. The change was done due to the low number of participants. Moreover, it was agreed upon to let the participants complete the survey on their own, as previous participants had no further issues or questions arising when completing the survey.

Participants of the study were invited to read the first page of the questionnaire explaining the purpose of this research and the difference between the Spanish and English version. They were further asked to select a language at the top right of the questionnaire. Thereafter, they could read the informed consent and if participants agreed to continue with the questionnaire, they were provided with the above-mentioned questions regarding their demographics. Thereafter, the questionnaire stated questions about their familiarity with the chatbot. Once this was completed, the participants had to answer the scale regarding their decision-making style. As a succeeding step, the researcher explained briefly how the

interaction with the conversational agent should go. Additionally, in the online session, the researchers made the participants aware that the researcher was going to stay in the session in case of questions or troubles. Without supervision, Qualtrics presented participants with a screen in which the same information was written down. The information clarified that the importance was on the interaction itself rather than on the correctness or completion of the task. Participants could then interact with the 10 chatbots at their own pace. After each completion of a task, they had to answer the UMUX-Lite Questionnaire (Lewis et al., 2013) as well as the BotScale (Balaji & Borsci, 2019). Once all ten interactions were completed, the researcher thanked the participant and asked whether any questions were left unanswered.

2.5 Data Analysis

Adjustments and Normality Test

Statistical analyses were conducted using R Studio (R Core Team, 2020). As the present data is of ordinal nature the normality of the data would be tested. This is as Siegel (1957) proposes that in the majority of cases ordinal data needs to be analyzed using nonparametric tests. To test for normality, researchers used the Shapiro-Wilk Test of Normality. Mudholkar et al. (1995) suggest that if the result of the test is significant it is an indication that the data is normally distributed. The test was run using the "dplyr" package (Mailund, 2019). The Q-Q plots were used to visualize the distribution of the data and assess whether the data is normally distributed. Researchers used the "ggqqplot" function of the "ggpubr" package for R (Kassambara, 2020). Based on the results, researchers decided to conduct further analysis using nonparametric statistical tests.

Moreover, a manipulation check was performed through a Mann-Whitney U test to test that there was no significant difference between the individuals that completed the BotScale supervised versus the participants that completed it unsupervised.

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) with the R package "lavaan" was conducted (Rosseel, 2012). Borsci et al. (2021, under review) found an underlying five-factor structure CFA, therefore, this structure was used to test this model (Appendix A). The goodness of fit of the model is divided into multiple measures. The first measure, the Model Chi-Square, is used to assess the overall fit of the model. Significant p values are considered a good fit for the model. The authors Hutchinson and Olmos (1998), however, suggest that this measure is sample size-dependent, where only large sample sizes result in significant p-values. The second shortcoming of this index, as described by these authors, is that especially non-normal data results in non-significant p-values which ultimately leads to extreme numbers of rejection of models. Moreover, the comparative fit index will be reported. With this analysis, the five-factor model is compared to a null model. Values until CFI=.90 are considered as an index of moderate fit (Lai & Yoon, 2015). The Root Mean Square Error of Approximation (RMSEA) is an index that compares the model to a perfect baseline model. It indicates the absolute fit of the model. Values below RMSEA<0.05 are considered indexes of a good fit of the model (Hancock & Freeman, 2001). Moreover, the Standardized Root Mean Square Residual (SRMR) assesses the difference between the observed and expected correlation. Values below SRMR<0.7 are considered indications of a good fit (Pavlov et al., 2021). The last two indexes used are primarily known to help in model selection. The Akaike Information Criterion (AIC) is especially important when comparing models as it indicates the quality of the model tested in relation to the other model (Vrieze, 2012). Thus, the model indicating the lowest AIC value can be seen as the model with the best fit. The Bayesian Information Criterion (BIC) is, additionally, used as a criterion to select the most fitting model. For this value, lower values represent a better fit of the model. (Vrieze, 2012)

To come to further decisions, regarding a new model a closer look at the factor loadings of each item was taken. Factor loadings represent the effect of the factor on the item. As a general rule, factor loadings of >0.6 are seen as acceptable if the analysis is done on established items (Peterson, 2000). Moreover, the loadings of each factor in relation to the satisfaction construct were drawn using the "sempath" function in the "semplot" package (Epskamp, 2015).

Reliability Analysis

Cronbach's Alpha was calculated to assess the reliability, more specifically the internal consistency of the BotScale and the UMUX Lite, by using the Psych package (Revelle, 2011). Additionally, the quality of each item was analyzed through the calculation of an item-total correlation. An index value below 0.3 demonstrates that the item does not correlate with the overall scale (Hwan, 2000).

Correlation Analysis

A Kendall's Tau test was performed to test the correlation between the BotScale and UMUX-Lite in line with the second hypothesis. The researchers employed the "kendall" package (McLeod, 2015).

Moreover, to explore the psychometric properties of the Spanish translation, first reliability coefficients are computed for both the original and the translated version. This was done not only on the overall scales but also per factor. Moreover, to see whether this Spanish version correlates with the properties of the English version a Kendall's Tau, a non-parametric correlation analysis was performed.

Finally, the median of the five different decision-making styles was calculated. Medians were used since the scale employs a Likert Scale (Sullivan & Artino, 2013). Additionally, the frequency of the style in the population was calculated in percentages. To test the relationship between the level of satisfaction and decision-making behaviour, researchers conducted a Kruskal-Wallis. With this test, the researchers can determine whether there is a significant difference in satisfaction levels between the different decision-making styles (McKight, 2010). The test was run using the "dplyr" package (Mailund, 2019)

3. Results

3.1 Descriptive Statistics

The medians of satisfaction as measured by the BotScale and the UMUX Lite were calculated, thereby the use of a Likert scale was accounted for (Boone & Boone, 2012).

Table 1

Median Satisfaction Levels

Questionnaire	Median	Standard Deviation
Item 1	4	0.544
Item 2	4	0.667
Item 3	4	0.222
Item 4	3.5	0.278
Item 5	3	0.322
Item 6	3.5	0.489
Item 7	4	0.177
Item 8	3	0.222
Item 9	4	0.222
Item 10	4	0.678
Item 11	4	0.4

Item 12	3.5	0.678
Item 13	4	0.678
Item 14	2	0.933
Item 15	4	0.456
BotScale Total	4	0.772
UMUX Lite Item 1	4	0.632
UMUX Lite Item 2	4	0.539
UMUX Lite Total	4	0.599

3.2 Normality Test and Data Manipulation

To answer the question of whether the data of the BotScale is normally distributed, we calculated two statistical tests. Firstly, the results of the Quantile-Quantile (Q-Q) Plot are shown in Figure 1.

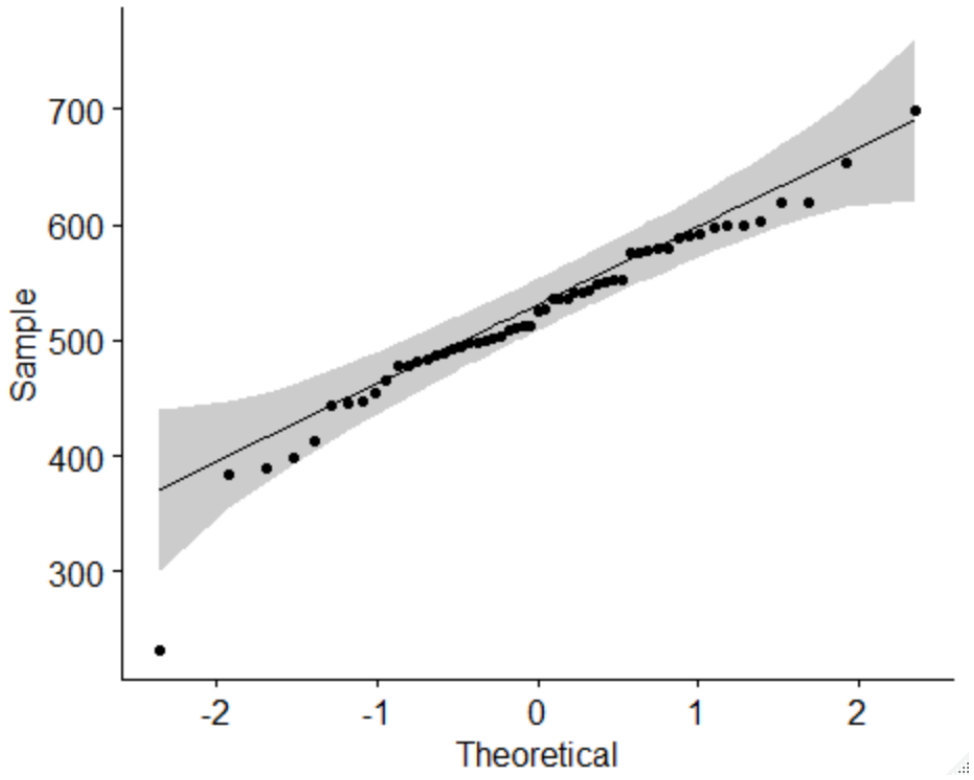

Figure 1 Q-Q Plot for Satisfaction Levels

Additionally, a Shapiro-Wilk normality test was run on the overall satisfaction levels, to see whether the variable is normally distributed. With this test, the sample distribution is compared to a normal distribution. The results of this test, W=0.950, p=0.023, reject the hypothesis that the data is normally distributed.

Furthermore, the results of a Mann-Whitney U test show that there is not a significant difference between the group that completed the survey supervised and the group that did it unsupervised (U=151.5, p=.386).

3.3 Confirmatory Factor Analysis

To answer the first research question of whether the previously found five-factor model can be confirmed a confirmatory factor analysis was performed.

Overall, the results of the goodness of fit of model 1 are ambiguous but mainly suggest that this model is unacceptable (CFI=.899, RMSEA=.154, SRMR=.06, AIC=4602.691, BIC=4678.970), Thus, the results suggest that further analysis on different models should be conducted.

Table 2

Goodness of fit of Model 1 for Satisfaction (N=53).

Model	X^2	Df	p	CFI	RMSEA	SRMR	AIC	BIC
Model 1- Five-Factor Model	177.4	82	.001	.899	.145	.063	4602.6	4678.9

Moreover, the modification index was calculated on the first model, to test whether the model can be improved using covariances. In that sense, it indicates whether adding a path in the model could improve the fit of the model. Results suggest that an additional link between Item 6 and Item 8 might improve the model. After running a further CFA adding this link, the results increased slightly, indicating a better fit for model 2 (SRMR=0.063, RMSEA=0.136, CFI=0.912, AIC=4591.103, BIC=4669.838).

Table 3

Goodness of fit of Model 2 for Satisfaction (N=53).

Model	X^2	Df	p	CFI	RMSEA	SRMR	AIC	BIC
Model 2 - Five Factor Model with covariance Items 6 and 8	163.8	81	.001	.912	.136	.063	4591.1	4669.8

Moreover, the unstandardized and standardized factor loadings of each item are presented in Table 4. The results of the standardized factor loadings indicate that all factor loadings, except for the factor loading for Item 8, are above 0.6. Since the overall fit of the model did not yield clear results, the low factor loading might indicate that item 8 can be removed from the questionnaire. To test whether model 3 would improve, a third analysis on a new model was run. Model 3 differs from the first five-factor model (Model 1), as Item 8 is removed from the list of items.

Table 4

Standardized Factor Loadings for Five-Factor Confirmatory Factor Model

Items	**F1**	**F2**	**F3**	**F4**	**F5**
Item 1	0.924				
Item 2	0.972				
Item 3		0.941			
Item 4		0.792			
Item 5		0.638			
Item 6		0.859			
Item 7		0.840			
Item 8		0.509			
Item 9		0.912			
Item 10			0.933		
Item 11			0.944		

Item 12	0.884		
Item 13	0.930		
Item 14		1	
Item 15			1

The third model, the five-factor model without Item 8, shows improved indexes of fit (CFI=0.938, RMSEA=.124, SRMR=.048, AIC=4272.763, BIC=4345.027). The value of the Root Mean Square Error of Approximation (RMSEA), however, is still higher than the ideal value of RMSEA<0.06. The results of the CFI have increased and show a moderate fit of the model (Table 4).

Table 5

Goodness of fit of Model 3 for Satisfaction (N=53)

Model	**X^2**	**Df**	**P**	**CFI**	**RMSEA**	**SRMR**	**AIC**	**BIC**
Model 3 -Five-Factor Model Without Item 8	123.3	69	.001	.938	.120	.048	4272.7	4345.0

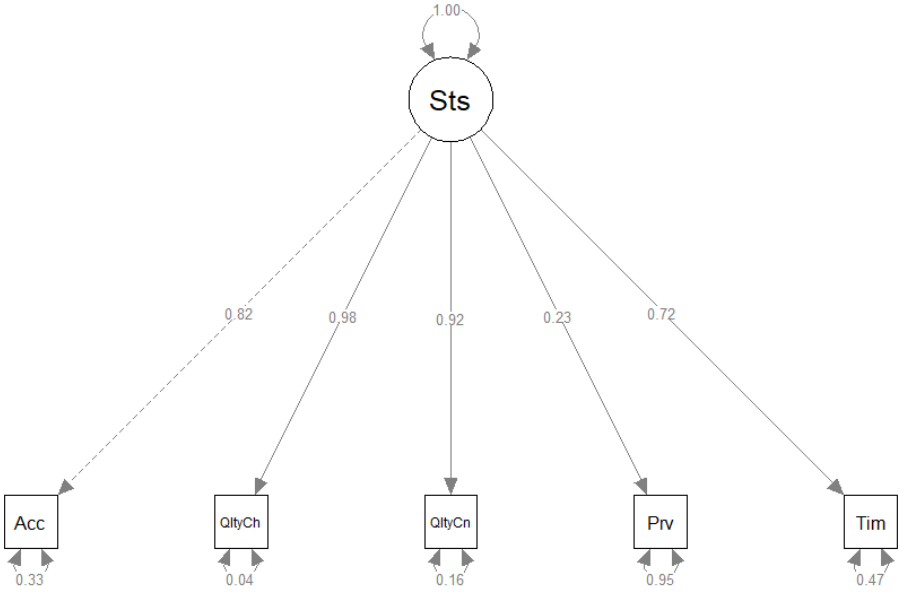

Figure 2 Factor Loading for the five-factor model

To better understand the relationship between the items, a visual representation of the loadings of the factors was drawn. The five factors displayed are the perceived accessibility of the chatbot (Acc); the perceived quality of the chatbot functions (QltCh); the perceived quality of conversation and information provided (QltCn); Perceived privacy and security (Prv) and Time and Response (Tim) (see also Appendix A). As the illustration shows, the fourth factor, privacy, has a low factor loading and was therefore removed. It was therefore determined that a fourth analysis would be run on a new model (model 4) consisting of four factors only. This model, model 4, based on four factors shows improved indexes (CFI=.943, RMSEA=.122, SRMR=.046, AIC=3881.536, BIC=3943.764). In this model, the SRMR value decreased while the CFI increased, which indicates both a moderate to a good fit of the model. The AIC value is significantly lower than in the previous models, indicating that this fourteen-item model has the best fit. This also accounts for the value of the BIC illustrating a good model. Lastly, the RMSEA value is still high in comparison to the recommended value of <.06 (Table 15). Overall, as this model displays the best indexed for the goodness of fit, this model will be used for further analysis. This new model consists of 14 items, and four factors namely the perceived accessibility of the chatbot (Acc); the perceived quality

of the chatbot functions (QltCh); the perceived quality of conversation and information provided (QltCn) and Time and Response (Tim). An illustration of the new model is represented in figure 3 (see also Appendix G). Hereafter, this new model will be referred to as the BotScale 14.

Table 6

Goodness of fit of Model 4 for Satisfaction (N=53)

Model	X^2	Df	p	CFI	RMSEA	SRMR	AIC	BIC
Model 4 - Four Factor Model	109.2	60	.001	.943	.122	.046	3881.5	3943.7

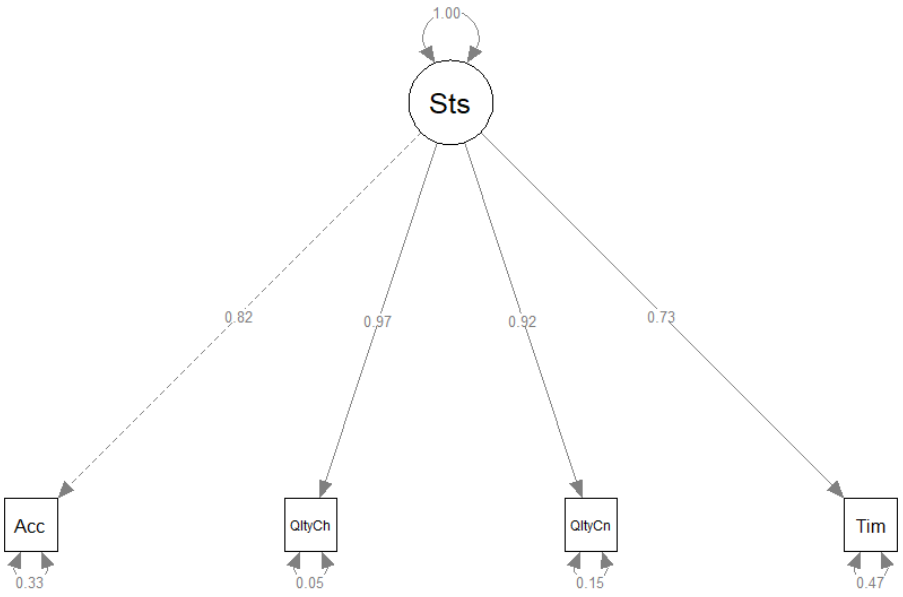

Figure 3 Factor Loading for the four-factor model

3.4 Reliability Analysis

To assess the quality of the items in the questionnaire a reliability analysis was conducted. Results are illustrated in Table 7. The value of r.drop indicates the total correlation of the scale without this item. If the value is low (<0.3), such as item 14, it is an indication that this item does not correlate with the overall scale. The Cronbach's alpha value of the 14-items questionnaire, the BotScale 14, was calculated and resulted in a high-reliability value of $\propto$=0.97, indicating a good internal consistency of the questionnaire. Furthermore, the Cronbach's alpha value for the UMUX Lite was calculated and the results indicate a high-reliability value of $\propto$=0.853.

Table 7

Item total correlations

Item	r.drop
Item 1	0.78
Item 2	0.83
Item 3	0.89
Item 4	0.77
Item 5	0.64
Item 6	0.85
Item 7	0.81
Item 8	0.63
Item 9	0.86

Item 10	0.88
Item 11	0.88
Item 12	0.79
Item 13	0.88
Item 14	**0.24**
Item 15	0.69

3.5 Correlation Analysis

3.5.1 Relationship between the BotScale14 and the UMUX Lite

A Kendall's tau, non-parametric correlation analysis was run to determine the relationship between the results of the BotScale14 and the UMUX Lite questionnaire. The results indicate that there is a significant positive correlation between these two scales (τ_b=0.69, p<0.001).

The Cronbach's Alpha of each Scale was calculated. The English version ($\propto$=0.94) and the Spanish version of the BotScale ($\propto$=0.94) results perfectly aligned, showing very good reliability for the two versions. Results suggested further a significant positive correlation between the Spanish and the English version of the scale (τ_b=0.842, p=0.007).

3.5.3 Relationship between Decision Making Styles and Satisfaction Levels

To test whether decision-making styles influence satisfaction levels one statistical test was computed. We performed descriptive statistics for the decision-making style (Table 8) and a Kruskal Wallis Test to observe whether decision-making styles influence satisfaction levels. This test showed that the Decision-Making Styles did not significantly influence satisfaction levels measured by the BotScale

($H(2)=6.026$, $p=0.19$). Additionally, we performed a further Kruskal Wallis Test on the satisfaction levels results by the UMUX Lite questionnaire, results showed that decision-making style did not significantly affect satisfaction levels ($H(4)=4.961$, $p=0.29$).

Table 8

Medians of satisfaction score for each decision-making style

	Intuitive (n=15)	**Dependent (n=9)**	**Avoidant (n=7)**	**Spontaneous (n=1)**	**Rational (n=23)**
Average satisfaction score on the BotScale	4	3.5	2.5	3	4
Frequency in the sample	27.27%	16.36%	12.73%	1.82%	41.82%

4. Discussion

4.1 Recapitulation and Implications of the present study

4.1.1 Psychometric Properties

The present research aimed at analyzing and confirming the psychometric properties of a newly developed scale for measuring satisfaction scores in the interaction with chatbots. The first research question thus was "*Can the factorial structure of the BotScale found in previous exploratory analyses be confirmed?"*. In that sense, it wanted to verify the scale developed by Borsci et al. (2021) for measuring the satisfaction with the interaction with chatbots. The data suggested that the initial model of five factors could be further reduced and optimised in terms of the number of items and factors. The best indexes of fit were yielded with a model, the BotScale 14, that is based on four underlying factors. The factors are perceived accessibility to chatbot functions, perceived quality of chatbot functions, perceived quality of conversation and information provided and time response (Appendix H). The first factor covers questions regarding whether it was easy to locate the chatbot. The perceived quality of chatbot functions asks the user whether the chatbot met the expectations based on general functions such as the context, conversation, and difficult situations. The fourth factor, namely the perceived quality of conversation, covers questions regarding the information received. The last factor, time response, asks the user whether the waiting time for a response was appropriate. Moreover, a four-factor structure was found in previous studies conducted by van den Bos and Borsci (2021) and, thus confirming the results from this previous study.

While looking at the results, however, none of the models tested in this analysis displays perfect or good indexes for all measurements. More specifically, the overall results of the confirmatory factor analysis showed that the value of the root mean square error of approximation (RMSEA) were not adequate for any of the models. A study by Kenny et al. (2015) indicated that when dealing with a small number of degrees of freedom, in the specific study up to 150 degrees of freedom, the results of this value

often indicate poorly fitting models even when this is not the case. It therefore can be cautiously concluded that more data are needed to further confirm the solution with 4 factors.

The results suggest that the BotScale correlates with the UMUX-Lite scale. Thus, being in accordance with the second research question, *"Do the results from the BotScale correlate with the results of the UMUX Lite?"*. The results are, furthermore, in line with previous results (Borsci et al., 2021, under review) that proposes a correlation between BotScale and standardised satisfaction scales. Nevertheless, further data should be collected. Moreover, the overall reliability of the BotScale14 suggested a robust construct behind this scale (Tavakol, & Dennick, 2011).

4.1.3 Spanish version of the scale

The third research question raised the question whether "*the Spanish translation of the BotScale present similar psychometric properties to the original version?"*. The data suggests that the Spanish version of the scale maintains the psychometric properties of the original scale. These results are promising as the translation of a validated scale has the main advantage of offering the possibility to gather data in a cross-cultural setting (Yu et al., 2004). Thus, by using a single and coherent measurement, the results can easily be compared in different populations. Additionally, a correct translation guarantees that individuals are able to answer the questionnaire in their native language. As Harzing (2005) proposes, differences in response patterns can be seen when comparing answers to the same questionnaire in different languages. Thus, he suggests that researchers should ask participants to answer questionnaires in their native language to ensure that researchers capture the true nature of the participants' thoughts and ideas towards the topic under research.

4.1.4 Satisfaction and Decision-Making Styles

Previous research focused on the relationship between satisfaction and decision-making in the context of chatbots and identified significant relationships among these concepts. Hence, the last research question proposed was whether *decision-making decision-making style of an individual influence the level*

of satisfaction in users of chatbots?". The results of the present study, however, are not in line with this research question and can not confirm this relationship. Hence, the satisfaction levels resulting from the BotScale14 are not affected by decision-making style. An effect of decision-making styles on satisfaction, as measured by the UMUX-Lite, could also not be found. The main idea supporting the hypothesis that decision-making styles influence satisfaction levels in chatbot interactions is that chatbots help in the decision-making process (De Vreede, Raghavan, & De Vreede, 2021). Previous research found that decision-making was especially significant in the area of consumer experience and that they tended to influence satisfaction levels. A significant result would have supported the idea that chatbots could be tailored, based on the preferred decision-making style of the user, to increase the level of satisfaction (Kazeminia et al., 2019). Thus, as this effect could not be found in this research, further research might be necessary to find a different variable that does significantly influence satisfaction levels. Afterwards, research on the influence of tailored chatbots on satisfaction levels can be done.

One specific hypothesis from a previous study of Ciovati (2020) was that a rational decision-making style would have a negative correlation with satisfaction level. Nevertheless, this can not be confirmed in this study. Possible reasons for this result can be the nature of the chatbots as they did not offer long interactions or many possible outcomes. In that sense people that rationally make decisions might have had the feeling that they were presented with all the information needed to come to a rational decision which resulted in overall high satisfaction levels (Cheek & Schwartz, 2016). Additionally, since the participants were not tested for the correctness of their completion of the task, individuals might not have felt pressured to explore alternatives than what was presented to them. Additionally, a second explanation for the difference in results might have been due to the different nature of the satisfaction scales used. While Ciovati (2020) employed the Questionnaire for User Interface Satisfaction (Chin, 1988), which focuses on satisfaction with the interface, this research employed the BotScale. The BotScale has a wider range of topics as it not only covers the interface and the functions but also the accessibility of the chatbot, the quality of the conversation, the time response.

4.2 Limitations and Future Research

This study presented three limitations. The first limitation that is accounted for by the current pandemic situation is that the study had to be held online. In that sense, participants had different experiences in chatbot interactions, with some chatbots malfunctioning or being under construction for example. This could have resulted in certain biases in the results. With this in mind, it is recommended to conduct further research under more controlled conditions.

This research focused on finding an underlying correlation between satisfaction and decision-making styles while the results were non-significant or low, not adding the personalization might account for these results. In this regard, a previous study by Ciovati (2020) found results after implementing and presenting the participants with personalized chatbots. It is therefore recommended to conduct further research on more interactive and tailored chatbots.

Thirdly, participants were asked to complete one task only. The reason behind this decision was that participants were asked to interact with 10 different chatbots to enable gathering more data of different chatbot interactions. To prevent participants from experiencing cognitive overload and therefore not finishing the study interactions were kept short. While the researchers tried to develop them in a way that enabled participants to equally interact with each chatbot overall interactions were short. It is therefore of value to research whether satisfaction levels change with longer interactions. Mittal et al. (2001) support this idea and suggest that overall satisfaction in the context of consumer experience can change over time. Furthermore, they argue that a longitudinal study of satisfaction is able to maximize satisfaction levels which ultimately might lead to different results than found in this study.

Lastly, further research is needed to test whether the results from this confirmatory factor analysis can be replicated and if the results from the Spanish version can be translated into larger samples for example.

4.3 Conclusion

As the global use of chatbots increases so does the need for a questionnaire to assess the quality of the chatbot as experienced by the user. Thus, the present study suggested a new structure for a recently developed scale to assess satisfaction with chatbots. This is especially crucial as it can be seen as a standardized and validated questionnaire that was tailored specifically for the measurement of satisfaction in the interaction with chatbots. Moreover, by showing evidence for a successful translation, this questionnaire enables a different population, namely a Spanish speaking population, to take part in research on chatbots. An influence of decision-making styles on satisfaction could not be found.

References

Alacreu-Crespo, A., Fuentes, M. C., Abad-Tortosa, D., Cano-Lopez, I., González, E., & Serrano, M. Á. (2019). Spanish validation of General Decision-Making Style scale: Sex invariance, sex differences and relationships with personality and coping styles. *Judgment and Decision Making*, *14*(6), 739. Retrieved from: https://www.researchgate.net/publication/337679701_Spanish_validation_of_General_Decision-Making_Style_scale_Sex_invariance_sex_differences_and_relationships_with_personality_and_coping_styles

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*. https://doi.org/10.1007/s12525-020-00414-7

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. Journal of the American Society for Information Science and Technology, 61(5), 859-868.https://doi.org/10.1002/asi.21300

Allan, J., Carterette, B., & Lewis, J. (2005, August). When will information retrieval be" good enough"?. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 433-440).https://doi.org/10.1145/1076034.1076109

Alavi, S. A., Rezaei, S., Valaei, N., & Wan Ismail, W. K. (2016). Examining shopping mall consumer decision-making styles, satisfaction and purchase intention. *The International Review of Retail, Distribution and Consumer Research, 26*(3), 272-303. https://doi.org/10.1080/09593969.2015.1096808

Andersen, K. E., Köslich, S., Pedersen, B. K. M. K., Weigelin, B. C., & Jensen, L. C. (2017). Do we blindly trust self-driving cars. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 67–68). New York, NY, USA: IEEE Computer Society. https://doi.org/10.1145/3029798.3038428

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

Armstrong, C. S., Banerjee, S., & Corona, C. (2013). Factor-loading uncertainty and expected returns. *The Review of Financial Studies*, *26*(1), 158-207. https://doi.org/10.1093/rfs/hhs102

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, *54*, 101473. https://doi.org/10.1016/j.tele.2020.101473

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

Banville, D., Desrosiers, P., & Genet-Volet, Y. (2000). Translating questionnaires and inventories using a cross-cultural translation technique. *Journal of teaching in physical education, 19*(3), 374-387. DOI: https://doi.org/10.1123/jtpe.19.3.374

Baroudi, J. J., & Orlikowski, W. J. (1988). A Short-Form Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use. Journal of Management Information Systems, 4(4), 44–59. doi:10.1080/07421222.1988.11517807

Berisha, G., Pula, J. S., & Krasniqi, B. (2018). Convergent validity of two decision making style measures. Journal of Dynamic Decision Making, 4, 1-1. DOI: https://doi.org/10.11588/jddm.2018.1.43102

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies*, *55*(4), 533–552. https://doi.org/10.1006/ijhc.2001.0483

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM* (Vol. 9, No. 1, pp. 1-4). Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.460.6252&rep=rep1&type=pdf

Bevan, N., Carter, J., & Harker, S. (2015). ISO 9241-11 Revised: What Have We Learnt About Usability Since 1998? *Human-Computer Interaction: Design and Evaluation, 9169,* 143–151. https://doi.org/10.1007/978-3-319-20901-2_13

Bieliauskas, S., & Schreiber, A. (2017). A Conversational User Interface for Software Visualization. In *Proceedings - 2017 IEEE Working Conference on Software Visualization, VISSOFT 2017* (pp. 139–143). IEEE. https://doi.org/10.1109/VISSOFT.2017.21

Bologna, C., De Rosa, A. C., De Vivo, A., Gaeta, M., Sansonetti, G., & Viserta, V. (2013). Personality-based recommendation in E-commerce. CEUR Workshop Proceedings, 997.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of extension, 50*(2), 1-5. Retrieved from https://www.researchgate.net/profile/Mahdi_Safarpour/post/what_is_a_logistic_regression_analysis/attachment/59d622fb79197b8077981513/AS:304626539139073@1449640034657/download/Likert+Scale+vs+Likert+Item+%28Good+Source%29.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., (2021,under review). The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *International conference on internet science* (pp. 377-392). Springer, Cham 10.1007/978-3-319-70284-1_30

Cheek, N. N., & Schwartz, B. (2016). On the meaning and measurement of maximization. *Judgment and Decision Making, 11*(2), 126–146. Retrieved from https://works.swarthmore.edu/fac-psychology/929

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 213-218).https://doi.org/10.1145/57167.57203

Chung, M., Ko, E., Joung, H. and Kim, S.J. (2018), "Chatbot e-service and customer satisfaction regarding luxury brands", *Journal of Business Research, Vol. 117,* pp. 587-595. https://doi.org/10.1016/j.jbusres.2018.10.004

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory*. Odessa, FL: Psychological assessment resources. Retrieved from: https://www.researchgate.net/profile/Paul-Costa/publication/247728869_Persons_Places_and_Personality_Career_Assessment_Using_the_Revised_NEO_Personality_Inventory/links/556c4f9008aec2268303bd9d/Persons-Places-and-Personality-Career-Assessment-Using-the-Revised-NEO-Personality-Inventory.pdf

Ciovati, A. (2020). Personalized Dyadic Chatbot Conversations: The influence of human and chatbot personality on customer satisfaction within the e-commerce domain. (Master thesis). Delft University of Technology, Delft, Netherlands.

De Vreede, T., Raghavan, M., & De Vreede, G. J. (2021). Design Foundations for AI Assisted Decision Making: A Self Determination Theory Approach. In Proceedings of the 54th Hawaii International Conference on System Sciences (p. 166). http://hdl.handle.net/10125/70630

Epskamp, S. (2015). semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling: a multidisciplinary journal*, *22*(3), 474-483. https://doi.org/10.1080/10705511.2014.93784

Eren, B. A. (2021). Determinants of customer satisfaction in chatbot use: evidence from a banking application in Turkey. *International Journal of Bank Marketing.* DOI 10.1108/IJBM-02-2020-0056

Feine, J., Morana, S., & Gnewuch, U. (2019). Measuring service encounter satisfaction with customer service chatbots using sentiment analysis. 14. Internationale Tagung Wirtschaftsinformatik (WI2019), pp 1115–1129

Ferreira, J. M., Acuna, S. T., Dieste, O., Vegas, S., Santos, A., Rodriguez, F., & Juristo, N. (2020). Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction. *Information and Software Technology*, *117*, 106195.https://doi.org/10.1016/j.infsof.2019.106195

Fischer, S., Soyez, K., & Gurtner, S. (2015). Adapting Scott and Bruce's General Decision-Making Style Inventory to Patient Decision Making in Provider Choice. *Medical Decision Making*, *35*(4), 525–532. https://doi.org/10.1177/0272989X15575518

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

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 345-352).https://doi.org/10.1145/332040.332455

Girard, A. J., Reeve, C. L., & Bonaccio, S. (2016). Assessing decision-making style in French-speaking populations: Translation and validation of the general decision-making style questionnaire. *European Review of Applied Psychology*, *66*(6), 325–333. https://doi.org/10.1016/j.erap.2016.08.001

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) 2017. 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

Gupta, A., Hathwar, D., & Vijayakumar, A. (2020). Introduction to AI Chatbots. Int. J. Eng. Res. Technol, 9(7). Retrieved from: https://www.researchgate.net/profile/Aishwarya-Gupta-17/publication/344895276_Introduction_to_AI_Chatbots/links/5f979f8392851c14bceab8d3/Introduction-to-AI-Chatbots.pdf

Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement, 61*(5), 741-758. https://doi.org/10.1177/00131640121971491

Hassan, H. M., & Galal-Edeen, G. H. (2017). From usability to user experience. In *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)* (pp. 216-222). IEEE. DOI: 10.1109/ICIIBMS.2017.8279761

Harzing, A.-W. (2005). Does the Use of English-language Questionnaires in Cross-national Research Obscure National Differences? *International Journal of Cross Cultural Management, 5(*2), 213–224. https://doi.org/10.1177/1470595805054494

Häubl, G., & Trifts, V. (2000). Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science, 19*(1), 4–21. doi:10.1287/mksc.19.1.4.15178

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods, 7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005). Freudbot: An investigation of chatbot technology in distance education. In *Proceedings of ED-MEDIA 2005--World Conference on Educational Multimedia, Hypermedia & Telecommunications (pp. 3913-3918).* Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/primary/p/20691/.

Helton, W. S. (2004). Validation of a short stress state questionnaire. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 11, pp. 1238-1242). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/154193120404801107

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior,* 49, 245-250. https://doi.org/10.1016/j.chb.2015.02.026

Holtgraves, T. M., Ross, S. J., Weywadt, C. R., & Han, T. L. (2007). Perceiving artificial social agents. *Computers in human behavior, 23*(5), 2163-2174. https://doi.org/10.1016/j.chb.2006.02.017

Hornbæk, K., & Law, E. L. C. (2007). Meta-analysis of correlations among usability measures. *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 617-626).https://doi.org/10.1145/1240624.1240722

Hoyle, R. H. (2000). Confirmatory factor analysis. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 465-497). Academic Press. https://doi.org/10.1016/B978-012691360-6/50017-3

Hutchinson, S. R., & Olmos, A. (1998). *Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. Structural Equation Modeling: A Multidisciplinary Journal, 5(4), 344–364.* doi:10.1080/10705519809540111

Hwang, I. H. (2000). The Usability of Item-Total Correlation as the Index of Item Discrimination. *Korean Journal of Medical Education*, 12(1), 45–51. Retrieved from https://www.koreamed.org/SearchBasic.php?RID=2306929

Io, H. N., & Lee, C. B. (2017). Chatbots and conversational agents: A bibliometric analysis. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 215-219). IEEE. https://doi.org/10.1109/IEEM.2017.8289883

ISO 9241-11. (2018) - Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. (n.d.). Retrieved April 4, 2019, from https://www.iso.org/standard/63500.html

Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. *Communications of the ACM*, *26*(10), 785-793. https://doi.org/10.1145/358413.358430

John, O. P. (1989). *Towards a Taxonomy of Personality Descriptors. Personality Psychology*, *261–271*. doi:10.1007/978-1-4684-0634-4_20

Joo, S. (2010). How are usability elements-efficiency, effectiveness, and satisfaction-correlated with each other in the context of digital libraries?. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1-2. https://doi.org/10.1002/meet.14504701323

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. *Sociological Methods & Research, 44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Khan R., Das A. (2018) Introduction to Chatbots. In: *Build Better Chatbots.* Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-3111-1_1

Kondo, Y. (2001). Customer satisfaction: how can I measure it?. Total Quality Management, 12(7-8), 867-872. https://doi.org/10.1080/09544120100000009

Lai, M. H., & Yoon, M. (2015). A modified comparative fit index for factorial invariance studies. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 236-248.https://doi.org/10.1080/10705511.2014.935928

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research,* 39(2 SPEC. ISS.), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lewis, J. R. (1995). *IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7(1), 57–78.* doi:10.1080/10447319509526110

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099-2102). DOI: 10.1145/2470654.2481287

Loo, R. (2000). A psychometric evaluation of the general decision-making style inventory. Personality and individual differences, 29(5), 895-905. https://doi.org/10.1016/S0191-8869(99)00241-X

Kaptein, M. C., Markopoulos, P., De Ruyter, B., & Aarts, E. (2010). Persuasion in ambient intelligence. *Journal of Ambient Intelligence and Humanized Computing, 1*(1), 43-56. DOI 10.1007/s12652-009-0005-3

Kassambara, A., (2020). Package 'ggpubr'. Retrieved from https://mran.microsoft.com/snapshot/2017-04-22/web/packages/ggpubr/ggpubr.pdf

Mailund, T. (2019). Manipulating data frames: dplyr. In *R Data Science Quick Reference* (pp. 109-160). Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4894-2_7

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019, March). A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th*

International Conference on Information and Education Technology (pp. 111-119). https://doi.org/10.1145/3323771.3323824

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's" adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of personality and social psychology*, *49*(3), 710. 10.1037//0022-3514.49.3.710

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1-1.https://doi.org/10.1002/9780470479216.corpsy0491

McLeod, A. I. (2015). Package 'Kendall'. *R Software: London, UK*. Retrieved from https://cran.microsoft.com/snapshot/2014-12-09/web/packages/Kendall/Kendall.pdf

Mudholkar, G. S., Srivastava, D. K., & Thomas Lin, C. (1995). *Some p-variate adaptations of the shapiro-wilk test of normality. Communications in Statistics - Theory and Methods, 24(4), 953–985.* doi:10.1080/03610929508831533

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. The Journal of Abnormal and Social Psychology, 66(6),574–583. doi :10.1037/h0040291

Oliveira, R. D., Cherubini, M., & Oliver, N. (2013). Influence of personality on satisfaction with mobile phone services. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *20*(2), 1-23. https://doi.org/10.1145/2463579.2463581

Oliver, R.L. (2010). Satisfaction: A Behavioral Perspective on the Consumer: A Behavioral Perspective on the Consumer (2nd ed.). Routledge. https://doi.org/10.4324/9781315700892

Oulasvirta, A., & Hornbæk, K. (2016, May). Hci research as problem-solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4956-4967). doi.org/10.1145/2858036.2858283

Paikari, E., & Van Der Hoek, A. (2018). A framework for understanding chatbots and their future. In *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)* (pp. 13-16). IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/8445528/authors#authors

Parker, A. M., De Bruin, W. B., & Fischhoff, B. (2007). Maximizers versus satisficers: Decision-making styles, competence, and outcomes. *Judgment and Decision making, 2(*6), 342. Retrieved from https://www.proquest.com/openview/471c277673390ae4cfa10ade75f8ec6e/1?pq-origsite=gscholar&cbl=696407

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the standardized root mean squared residual (SRMR) to assess exact fit in structural equation models. *Educational and Psychological Measurement, 81*(1), 110-130. https://doi.org/10.1177/0013164420926231

Peterson, R.A. (2000) A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters* 11, 261–275 . https://doi.org/10.1023/A:1008191211004

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic Information Systems. *Business & Information Systems Engineering*, *61*(4), 523–533. https://doi.org/10.1007/s12599-019-00599-y

Polman, E. (2010). Why are maximizers less happy than satisficers? Because they maximize positive and negative outcomes. *Journal of Behavioral Decision Making*, *23*(2), 179-190.https://doi.org/10.1002/bdm.647

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (Eds.). (2004). Methods for Testing and Evaluating Survey Questionnaires. Wiley Series in Survey Methodology. doi:10.1002/0471654728

Quadrelli, S., Davoudi, M., Galíndez, F., & Colt, H. G. (2009). Reliability of a 25-item low-stakes multiple-choice assessment of bronchoscopic knowledge. *Chest, 135*(2), 315-321.https://doi.org/10.1378/chest.08-0867

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

Raffaldi, S., Iannello, P., Vittani, L., & Antonietti, A. (2012). Decision-Making Styles in the Workplace. *SAGE Open*, *2*(2), 215824401244808. https://doi.org/10.1177/2158244012448082

Revelle, W. (2011). An overview of the psych package. *Dep Psychol Northwest Univ*, *3*, 1-25.https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.7429&rep=rep1&type=pdf

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1-36. Retrieved from https://users.ugent.be/~yrosseel/lavaan/lavaanIntroduction.pdf

Sarsam, S. M., & Al-Samarraie, H. (2018). A First Look at the Effectiveness of Personality Dimensions in Promoting Users' Satisfaction With the System. *SAGE Open*, *8*(2), 215824401876912. https://doi.org/10.1177/2158244018769125

Sanny, L., Susastra, A., Roberts, C., & Yusramdaleni, R. (2020). The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia. *Management Science Letters*, *10*(6), 1225–1232. doi: 10.5267/j.msl.2019.11.036

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1609-1618). https://doi.org/10.1145/1518701.1518947

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement, 55*(5), 818-831. https://doi.org/10.1177/0013164495055005017

Shumanov, M., & Johnson, L. (2020). Making conversations with chatbots more personalized. *Computers in Human Behavior*, *117*(0747-5632), 106627. https://doi.org/10.1016/j.chb.2020.106627

Siegel, S. (1957). *Nonparametric Statistics. The American Statistician, 11(3), 13–19.* doi:10.1080/00031305.1957.10501091

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.

Soni, V. D. (2018). Prediction of Genuinity of News using advanced Machine Learning and Natural Language processing Algorithms. *International Journal of Innovative Research in Science Engineering and Technology, 7(*5), 6349-6354.DOI: 10.15680/IJIRSET.2018.0705232

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management, 28*(4), 503–516. https://doi.org/10.1016/0306-4573(92)90007-m

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education*, *5*(4), 541. doi: 10.4300/JGME-5-4-18

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

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, 2, 53. doi: 10.5116/ijme.4dfb.8dfd

Thong, J. Y. L., & Yap, C.-S. (1996). Information systems effectiveness: A user satisfaction approach. *Information Processing & Management*, *32*(5), 601–610. https://doi.org/10.1016/0306-4573(96)00004-0

Thunholm, P. (2004). Decision-making style: habit, style or both?. *Personality and individual differences*, *36*(4), 931-944. https://doi.org/10.1016/S0191-8869(03)00162-4

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2017). Here's What I Can Do. In Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems (pp. 1–10). New York, NY, USA: ACM. https://doi.org/10.1145/3160504.3160544

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, Netherlands.

Verma, N. and Rangnekar, S. (2015), "General decision making style: evidence from India", *South Asian Journal of Global Business Research,* Vol. 4 No. 1, pp. 85-109. https://doi.org/10.1108/SAJGBR-09-2013-0073

Vrieze, S. I. (2012). *Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods, 17(2), 228–243.* doi:10.1037/a0027127

Waldera, L., & Borsci, S. (2019). Development of a Preliminary Measurement Tool of User Satisfaction for Information-Retrieval Chatbots. (Bachelor Thesis). University of Twente, Enschede, Netherlands.

Walker, D. A. (2003). JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods, 2*(2), 26. DOI: 10.22237/jmasm/1067646360

Xu, A., Liu, Z., Guo, Y., Sinha, V.,& Akkiraju, R. (2017). "A new chatbot for customer service on social media.". In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3506– 3510. https://doi.org/10.1145/3025453.3025496

Yu, D. S. F., Lee, D. T. F., & Woo, J. (2004). Issues and Challenges of Instrument Translation. Western Journal of Nursing Research, 26(3), 307–320. https://doi.org/10.1177/0193945903260554

Zemčík T. (2019). A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*. doi: 10.12783/dtcse/aicae2019/31439

Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting Virtual Agents. *ACM Transactions on Interactive Intelligent Systems*, *9*(2-3), 1–36. https://doi.org/10.1145/3232077

Zumstein, D., & Hundertmark, S. (2018). Chatbots: an interactive technology for personalized communication and transaction. *IADIS International Journal on www/Internet*, *15*(1), 96-109. Retrieved from: http://www.iadisportal.org/ijwi/papers/2017151107.pdf

Appendices

Appendix A

BotScale

Chabot Satisfaction Scale (15 Items). The current version was tested with a five-point Likert scale from 1 ("Strongly Disagree") to 5 ("Strongly Agree")

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.

Appendix B

BotScale Spanish Translation

Chabot Satisfaction Scale (15 Items). Translated Version.

Factor	Item
1 - Perceived accessibility to chatbot functions	1. La función del chatbot fue fácilmente detectable.
	2. Fue fácil encontrar/localizar el chatbot.
2 - Perceived quality of chatbot functions	3. La comunicación con el chatbot fue clara.
	4. Me enteré inmediatamente de la información que me puede dar el chatbot.
	5. La interacción con el chatbot se sintió como una conversación en curso
	6. El chatbot fue capaz de realizar un seguimiento del contexto.
	7. El chatbot pudo hacer referencias al sitio web o al servicio cuando fue necesario.

	8. El chatbot podía manejar situaciones en las que la línea de conversación no estaba clara
	9. Las respuestas del chatbot fueron fáciles de entender.
3 - Perceived quality of conversation and information provided	10. Encuentro que el chatbot comprende lo que quiero y me ayuda a lograr mi objetivo.
	11. El chatbot me da la cantidad adecuada de información.
	12. El chatbot solo me da la información que necesito.
	13. Siento que las respuestas del chatbot fueron precisas.
4 - Perceived privacy and security	14. Creo que el chatbot me informa sobre posibles problemas de privacidad.
5 - Time response	15. Mi tiempo de espera para recibir una respuesta del chatbot fue breve.

Appendix C

UMUX Lite

The current version was tested with a five-point Likert scale from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”)

“This system’s capabilities meet my requirements”
“This system is easy to use.”

Appendix D

General Decision Making Style (GDMS) Scale

The current version was tested with a five-point Likert scale from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”)

1. When I make decisions, I tend to rely on my intuition (I)
2. I rarely make important decisions without consulting other people (D)
3. When I make a decision, it is more important for me to feel the decision is right than to have a rational reason for it(I)
4. I double check my information sources to be sure I have the right facts before making a decision (R)
5. I use the advice of other people in making my important decisions (D)
6. I put off making decision because thinking about them makes me uneasy (A)

7. I make decisions in a logical and systematic way (R)
8. When making decisions I do what feels natural at the moment (S)
9. I generally make snap decisions (S)
10. I like to have someone steer me in the right direction when I am faded with important decisions (D)
11. My decisions making requires careful thought (R)
12. When making a decision, I trust my inner feelings and reactions (I)
13. When making a decision, I consider various options in terms of a specific goal (R)
14. I avoid making important decisions until the pressure is on (A)
15. I often make impulsive decisions (S)
16. When making decisions, I rely upon my instincts (I)
17. I generally make decision that feel right to me (I)
18. I often need the assistance of other people when making important decisions (D)
19. I postpone decision making whenever possible (A)
20. I often put off making important decisions (A)
21. I often put off making important decisions (A)
22. If I have the support of other, it is easier for me to make important decisions (D)
23. I generally make important decision at the last minute (A)
24. I make quick decisions (S)

25. I usually have a rational basis for making decision (R)

Appendix E

Consent Form

Taking part in the study

I have read and understood the study information. I consent voluntarily to be a participant in this study and

understand that I can refuse to answer questions and I can withdraw from the study at any time, without

having to give a reason. I understand that taking part in the study involves me interacting with different

chatbots. The whole experiment will take about 60 minutes. I understand that for participating in the study

there are no known risks involved. I am at least 18 years old.

Use of the information in the study

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

Future use and reuse of the information by others

I understand that information I provide will be used for a bachelor thesis. I understand that before the

information is achieved it will be anonymized by removing name and other information that could track me back. I give permission for the filling out of the scales and demographics questionnaire that I provide to be archived in a safe data repository so it can be used for future research and learning. Contact Information for Questions about Your Rights as a Research Participant If you ever have any questions after this session has ended you can email me: s.m.kerwienlopez@student.utwente.nl and my supervisor can be reached at s.borsci@utwente.nl. If you have questions about your rights as a research participant or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-bms@utwente.nl

• I understand and agree to participate voluntarily

• No, I would like to end this session

Appendix F

Tasks

1 - University of Twente	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's at the Utwente in the Netherlands. You ask the Utwente chatbot

	what options for a scholarship are available.
2 - Amtrak	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.
3 - Lufthansa	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.
4 - Emirates	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 of September until the 9th of October to London for two persons.
5 - HDFC Bank	Perform the following task using the chatbot: You are new to online banking and would like to know what a SIP is.
6 - Ibenta	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.

7 - Benefit	Perform the following task using the chatbot: You are interested in buying brown mascara. Find out what options there are.
8 - Gol	Perform the following task using the chatbot: You want to know which destination GOL flies to, you are interested in national destinations in the southern area.
9 - Absolut Vodka	Perform the following task using the chatbot: You are interested in finding out where the Absolut is from.
10 - ChatBot	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.

Appendix G

R code

```
#libraries
library(tidyverse)
```

```
library(dplyr)
library(ggpubr)
library(haven)
library(readxl)
library(lavaan)
library(psych)
#normality tests
shapiro.test(ChatbotMay_17$SatisfactionTotal)
ggqqplot(ChatbotMay_17$SatisfactionTotal)
wilcox.test(DataSupervised$SatisfactionTotal, DataUnsupervised$SatisfactionTotal)
#confirmatory factor analysis model 1
model1 <-'Accessibility=~Item1+Item2
    QualityChatbot=~Item3+Item4+Item5+Item6+Item7+Item8+ Item9
    QualityConversation=~Item10+Item11+Item12+Item13
    Privacy=~Item14
    Time=~Item15'
fit<-cfa(model1, data=ChatbotMay_17)
summary(fit, fit.measures=TRUE,standardized=TRUE, ci=TRUE)
modindices(fit, minimum.value = 10, sort=TRUE)
inspect(fit, what="std")
model_cov<-'Accessibility=~Item1+Item2
      QualityChatbot=~Item3+Item4+Item5+ Item6 +Item7+ Item8 +Item9
      QualityConversation=~Item10+Item11+Item12+Item13
      Privacy=~Item14
      Time=~Item15
```

```
      Item6~~Item8'
fitcov <- cfa(model_cov, data=ChatbotMay_17)
summary(fitcov, ci=TRUE, standardized=TRUE, fit.measure=TRUE)
#confirmatory factor analysis model 2
model3 <-'Accessibility=~Item1+Item2
     QualityChatbot=~Item3+Item4+Item5+Item6+Item7+Item9
     QualityConversation=~Item10+Item11+Item12+Item13
     Privacy=~Item14
     Time=~Item15'
fit2 <-cfa(model3, data=ChatbotMay_17)
summary(fit2, ci=TRUE,standardized=TRUE, fit.measure=TRUE)
modindices(fit2, minimum.value = 10, sort=TRUE)
model4 <- 'Accessibility=~Item1+Item2
  QualityChatbot=~Item3+Item4+Item5+Item6+Item7+Item9
  QualityConversation=~Item10+Item11+Item12+Item13
  Time=~Item15'
fit4 <-cfa(model4, data=ChatbotMay_17)
summary(fit4, ci=TRUE,standardized=TRUE, fit.measure=TRUE)
#drawing the indexes
satisfaction <- 'Satisfaction=~Accessibility+QualityChatbot+QualityConversation+Privacy+Time'
fitsatisfaction <- cfa(satisfaction, data=ModelData)
semPaths(fitsatisfaction,whatLabels="stand",layout = "tree")
#reliability analysis
alpha(Chatbot_SurveyMay_17$SatisfactionOverall)
alpha(Chatbot_SurveyMay_17$SatisfactionTotal)
```

```
#correlation analysis
#1st UMUX
cor(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$UMUXOverall,
method="kendall")
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$UMUXOverall,
method="kendall")
#2nd translated version
cor(EnglishData$SatisfactionOverall, SpanishDataSet$SatisfactionOverall, method="kendall")
cor.test(Translation$English, Translation$Spanish, method="kendall")
alpha(EnglishData$SatisfactionOverall)
alpha(SpanishDataSetFINAL)
#3rd Decision Making Styles
median(Chatbot_SurveyMay_17$Intuitive)
median(Chatbot_SurveyMay_17$Avoidant)
median(Chatbot_SurveyMay_17$Rational)
median(Chatbot_SurveyMay_17$Spontaneous)
median(Chatbot_SurveyMay_17$Dependent)
kruskal.test(SatisfactionOverall~DMS, data=Chatbot_SurveyMay_17)
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$Intuitive,
method="kendall")
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$Dependent,
method="kendall")
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$Avoidant,
method="kendall")
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$Spontaneous,
```

```
method="kendall")
cor.test(Chatbot_SurveyMay_17$SatisfactionOverall, Chatbot_SurveyMay_17$Rational,
method="kendall")
```

Appendix H

BotScale 14

Chabot Satisfaction Scale (14 Items). The current version was tested with a five-point Likert scale from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”)

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 - Time response	14. My waiting time for a response from the chatbot was short.