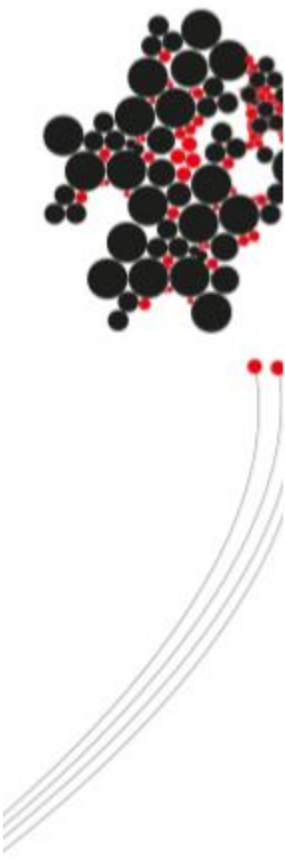
Assessing User Satisfaction with Information Retrieval Chatbots

Simone Wilmer – s2260425
Human Factors and Engineering Psychology
Faculty of Behavioural, Management, and Social Sciences
Department of Cognitive Psychology and Ergonomics

University of Twente



Supervision:
Dr. Simone Borsci
Prof. Dr. Frank Van der Velde

January 2021

UNIVERSITY
OF TWENTE.

Abstract

This study aims to develop an instrument that measures user satisfaction with chatbots, as such instruments are currently solely existing for vocal interaction experiences. The current study follows from previous research that established a preliminary 42-item questionnaire (USQ), aimed at measuring user satisfaction with chatbots. The additional value of this study lies in the revision and compression of the USQ. For this purpose, 48 participants interact with five chatbots and fill out the USQ after each interaction. A Principal Component Analysis is conducted, which leads to the establishment of a 24-item, four-factor USQ.

Next, it is examined whether the 24-item USQ consists of sufficient internal consistency. Reliability assessment of the 24-item USQ indicates that the overall USQ score, as well as the individual factors, consist of sufficient reliability. Also, the 24-item USQ's (concurrent) validity is examined, which is sufficient as well. This validity assessment is based on comparing scores on the 24-item USQ and its individual factors, with scores of an already existing questionnaire; the SUISQ-MR.

This study additionally focuses on possible confounding effects of chatbot-familiarity and geekism, which are examined with the support of non-linear regression analyses. Solely chatbot-familiarity has a significant upward-opening parabolic influence on the scores of the 24-item USQ.

Concludingly, this study confirms the results of previous research. Yet, it also provides additional insight in possible future research directions, advancing the ultimate goal; deploying a final revised version of the USQ that measures user satisfaction with chatbots.

Keywords: chatbots, familiarity, geekism, user satisfaction

Table of contents

1. Introduction

1.1 Human-computer interaction and chatbots

Chatbots are user interfaces that use natural language in their conversations to stimulate the interaction between human and technology. In contrast to voice-based interfaces, where users interact with the interface by directly speaking to the system, chatbots are textual conversational agents which provide service by means of conversation exchange via messages. Although chatbots have been around for a significant amount of time, they are currently gaining more popularity in our everyday life. Moreover, Piccolo, Mensio, and Alani (2018) state that besides voice-based interfaces like Apple's Siri, a rise in the popularity of textual chatbots occurred over the past years. It is even argued that chatbots will be the next main source of online information retrieval (Følstad & Brandtzæg, 2017).

In addition to online information retrieval, there are also other relevant benefits that come with deploying chatbots. For instance, Shawar and Atwell (2007) state that chatbots and conversational agents can serve a purpose that more conventional information retrieval systems, like search engines, often cannot provide. Namely, differently from search engines, chatbots may support users in information retrieval tasks by a natural process of conversation exchange. As a result, online communication could be perceived as more time efficient, allowing users to find information faster (Shawar & Atwell, 2007). Hence, deployment of chatbots in online communication may serve a significant purpose in more efficient, and satisfactory experiences (Følstad & Brandtzæg, 2017; Shawar & Atwell, 2007).

Since the use of chatbots is relatively new in online communication, interaction with chatbots may initially require high cognitive demands from its users (Hill, Ford, & Farreras, 2015). In addition, a high cognitive load is likely to result in more negative experiences with regard to the quality of chatbot interaction. This is mostly due to the novel environment users encounter when interacting with a chatbot rather than with a human being. Novel environments require users to rely on their history of familiar experiences and their intuition (Hill et al., 2015; Van Hooij, 2016). With regard to chatbots, this history of familiar experiences is lacking, which increases the difficulty to communicate with chatbots.

To summarize, chatbots may be useful for online information retrieval tasks. Yet, there are significant differences in content and quality between human-human online interaction and human-chatbot interaction (Følstad & Brandtzæg, 2017; Hill et al., 2015; Koopman & Schmettow, 2019; Mathur & Reichling, 2016). Namely, human-chatbot interaction is characterized by shorter messages and longer conversation durations, as opposed to human-human online interaction. Furthermore, Hill et al., (2015) state that human-

chatbot interaction differs from human-human interaction in the richness of vocabulary and usage of profanity. An important finding should be highlighted. Namely, although chatbots are not yet developed in a way that they can efficiently replace human beings, people are still willing to invest in the deployment of chatbots in online communication (Hill et al., 2015).

1.2 Conceptualizing and operationalizing user satisfaction with chatbots.

Before chatbots are utilized for information retrieval and communicational services, they should be tested on their usability by potential end-users (Følstad & Brandtzæg, 2017; Hill et al., 2015). Testing a chatbot's usability is relevant for gaining an understanding of how chatbots are perceived by end-users. A chatbot's usability is tested by means of users' direct interaction with chatbots while simultaneously being observed and questioned.

Usability is defined by ISO 9241-11 as the extent to which specific goals can be achieved with effectiveness, efficiency, and satisfaction by specified users, in specified contexts of use (ISO, 2018). Effectiveness is thereby defined as the accuracy and completeness of results and the level of achievement of users' goals or tasks (Frøkjær, Hertzum, & Hornbæk, 2000). In addition, Følstad and Brandtzæg (2017) state that a chatbot's ability to support conversational processes, while still providing useful output, correlates with the success of that chatbot as a natural language user interface. Whereas effectiveness is more goal-oriented, efficiency focuses on the effort needed for achieving results of a certain quality (Frøkjær, Hertzum, & Hornbæk, 2000). Whereas effectiveness and efficiency are performance measures (Bevan, 1995), satisfaction is dependent on users' preferences. User satisfaction concerns whether users like or dislike a system and can be measured by means of multiple tools, such as attitude rating scales and measurements of user preference (Frøkjær, Hertzum, & Hornbæk, 2000; Kirakowski & Corbett, 1993).

Increasing numbers of scientific research are focusing on how to improve chatbots to reach desirable effects regarding usability. For instance, Coperich, Cudney, and Nembhard (2017) examine how chatbots can be set up with the support of a human factors methodology. A human factors methodology involves comparing different chatbots on their results on usability tests. User satisfaction is thereby influenced by the quality as perceived by the user while interacting with a system like a chatbot (Coperich et al., 2017). Simultaneously, satisfaction with chatbots is influenced by perceptions on the other two factors (i.e. effectiveness and efficiency) (Bevan, 1995). Figure 1 illustrates a schematic overview of the factors influencing user satisfaction and an overall overview of the measurements of quality of use.

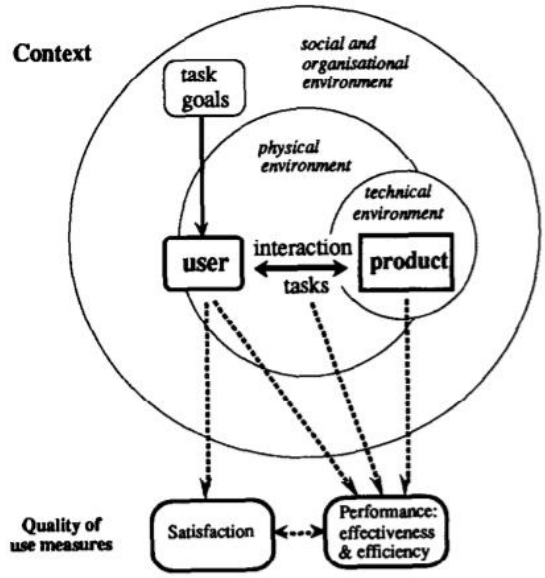


Figure 1. Measures of quality of use (Bevan, 1995).

Thus far, satisfaction was conceptualized and operationalized within the setting of usability testing by several researchers (Bevan, 1995; Coperich et al., 2017; Kirakowski & Corbett, 1993; Frøkjær et al., 2000). Next, it is important to establish a deeper understanding of user satisfaction and its possible influence on chatbot interaction, and vice versa. Firstly, Duijst (2017) introduced user satisfaction as a basic need within the overall user experience. User experience is defined by ISO 9241-210 as reactions and responses resulting from the use and/or anticipated use of a system (ISO, 2010). This is based on the idea that user satisfaction is influenced by the context of interaction, which implies that user satisfaction is dependent on to-be-performed tasks and contexts of communication (Kiseleva et al., 2016). Secondly, an important insight in user satisfaction is that it can rely on a variety of variables like trustworthiness (Böcker & Borsci, 2019), confidence (Deci & Ryan, 1985), and perceived competence (Macaranas, Antle, & Riecke, 2015). These variables are not only significant for understanding user satisfaction with chatbots, but also allow increased feasibility of satisfaction by considering human factors in the design process.

A more recent operationalization can be accounted for by Tariverdiyeva and Borsci (2019), who aimed for establishing this deeper understanding of user satisfaction and its influence on user perception regarding chatbots. Moreover, Tariverdiyeva and Borsci (2019) operationalized user perception of chatbots by establishing 18 key features (Appendix A). These 18 features illustrate the usability concepts that are important for a positive attitude towards a certain chatbot. Subsequently, these features have been compressed in a 14-feature version (Table 1). This version has been used as a framework for the development of a

preliminary 42-item questionnaire; the User Satisfaction Questionnaire (Appendix B), aimed at measuring user satisfaction with chatbots (Balaji & Borsci, 2019).

Table 1.

Revised features on user perception with chatbots (Balaji & Borsci, 2019)

	Key feature	Feature description
1	Ease of starting a conversation	How easy it is to start interacting with the chatbot
2	Accessibility	The ease with which the user can access the chatbot
3	Expectation setting	The extent to which the chatbot sets expectations for the interaction with an emphasis on what it can and cannot do
4	Communication effort	The ease with which the chatbot understands a range of user input
5	Ability to maintain themed discussion	The ability of the chatbot to maintain a conversational theme once introduced and keep track of context
6	Reference to service	The ability of the chatbot to make references to the relevant service
7	Perceived privacy	The extent to which the user feels the chatbot protects one's privacy
8	Recognition and facilitation of user's goal and intent	The ability of the chatbot to understand the user's intention and help them accomplish their goal
9	Relevance	The ability of the chatbot to provide information that is relevant and appropriate to the user's request
10	Maxim of quantity	The ability of the chatbot to respond in an informative way without adding too much information
11	Graceful breakdown	The ability of the chatbot to respond appropriately when it encounters a situation it cannot handle
12	Understandability	The ability of the chatbot to communicate clearly and in an easily understandable manner
13	Perceived credibility	The extent to which the user believes the chatbot's responses to be correct and reliable
14	Perceived speed	The ability of the chatbot to respond timely to user's requests

1.3 The current study

Until recently, satisfaction scales mainly assess user experiences with vocal interfaces, like Apple's Siri, rather than user perceptions with textual conversational interfaces like chatbots (Piccolo, 2018; Goossens, n.d.). Moreover, several instruments have been developed to assess the usability of vocal interaction experiences, like the MOS (Lewis, 2017), SASSI (Hone & Graham, 2000), and SUISQ (Lewis & Hardzinski, 2015). However, there are no similar instruments developed for chatbots. Hence, the current study is aimed at filling this research gap by developing an instrument for interactions with chatbots. More specifically, this instrument is ought to measure an overall experience with textual chatbots.

As a starting point, this study elaborates on the preliminary instrument (i.e. USQ) that was introduced by Balaji and Borsci (2019). The USQ should be further refined by means of validating and assessing the reliability of the scale and its subscales. Validation of the USQ will be supported by the addition of another questionnaire; the SUISQ-MR. The SUISQ-MR is a validated questionnaire (Lewis & Sauro, 2020) and it will be used to perform an external validation of the USQ. Moreover, the present study will also control for influences on USQ scores of 'familiarity with chatbots' and 'interest in technology or geekism'. Familiarity with technology has shown to be influential on several aspects of perceived usability, perceived self-efficacy, and performance (Fu & Gray, 2004; Tuch, Presslaber, Stöcklin, Opwis, & Bargas-Avila, 2012; Payne, Richardson, & Howes, 2000). Additionally to familiarity, also 'interest in technology or geekism' will be assessed on a possible influence on scores on the USQ, since it has shown to influence generic user experiences (Dehmel & Borsci, 2020; Van Hooij, 2016).

In sum, the current study will further refine the preliminary 42-item USQ by means of usability testing with a set of chatbots and potential end-users. The refinement of the USQ will lead to the proposal of a final and shortened version of the USQ. The final proposed questionnaire will be reduced in the number of items, only containing the items that have shown to be of most significant influence on measuring user satisfaction. Hence, the overall goal is to arrive at a refined USQ that measures an overall experience with textual chatbots.

This research goal will be met with the support of the following established research questions that extend from previously discussed literature:

"Q1: Does the current study confirm previously proposed factorial USQ-structures?"

"Q2: Do the results of the final proposed USQ correlate with the results of the SUISQ-MR?"

"Q3: Does familiarity with chatbots affect participants' answers to the USQ?"

"Q4: Does interest in technology (geekism) affect participants' answers to the USQ?"

2. Method

2.1 Participants

In total, a sample of 48 participants was recruited via the BMS Test Subject Pool system (SONA) and via the usage of convenience sampling. The sample consisted of 30 females and 18 males who ranged from ages 18 to 30 (Mean = 21.9, Standard Deviation = 2.55). The majority of participants were of Dutch (48.8%) or German nationality (35.4%), whereas other nationalities were observed as well, like Bulgarian (2.1%), Chinese (4.2%), Croatian (2.1%), Irish (2.1%), Lithuanian (2.1%), and Romanian (6.3%).
Participants tested five chatbots. Therefore, every participant filled in the USQ five times, leading to a total number of filled in USQs of 240. Some relevant exclusion criteria were used in this study. Namely, people under the age of eighteen and people who did not master the English language were not eligible to participate. Lastly, participants who did not fully agree to the informed consent, were excluded from this study.

2.2 Materials

The materials for this study include a Qualtrics survey, a researcher script (Appendix C), Google Hangouts, and twelve chatbots with two accompanying tasks per chatbot (Appendix C).

2.2.1 Qualtrics survey

The Qualtrics survey is based on previous research (Böcker & Borsci, 2019; Balaji & Borsci, 2019), and contains, among others, the preliminary USQ. The preliminary USQ contains 42 items and is set up with the support of the revised feature list that was introduced by Balaji and Borsci (2019) (Table 1). Moreover, every established feature was measured through three items in the preliminary USQ (Appendix B). The USQ items were displayed on a 5-point Likert scale (Appendix B), ranging from strongly disagree to strongly agree.
Additionally to the preliminary USQ, the Qualtrics survey also incorporates the SUISQ-MR (Speech User Interface Service Quality- Maximally Reduced). The SUISQ-MR is deployed to support the validation process of the USQ. The SUISQ-MR assesses Interactive Voice Response (IVR) utilizing 9 items that measure User Goal Orientation (item 1 and 2), Customer Service Behaviour (item 3 and 4), Speech Characteristics (item 5 and 6), and Verbosity (item 7, 8, 9) (Lewis & Hardzinski, 2015). The SUISQ-MR items were displayed on a 7-point Likert scale (Appendix B), ranging from strongly disagree to strongly agree. In addition, the SUISQ-MR was displayed with an additional ‘NA’ option, since not all items were applicable to each chatbot.

A third questionnaire is added in the Qualtrics survey that measures interest and enthusiasm in technology (geekism) (Schmettow & Drees, 2014). The Geekism-questionnaire consists of 15 items and was displayed on a 5-point Likert scale ranging from 'I totally disagree' to 'I totally agree' (Appendix B). In contrast to the USQ and SUISQ-MR, this questionnaire will only be filled out by partcipants once, at the end of the survey. Beside these questionairres, also generic questions, like task difficulty questions, demographical questions, as well as an informed consent form, are incorporated in the first part of the Qualtrics survey. The item that measures chatbot familiarity ('How familiar are you with chatbots/and or other conversational interfaces?') is part of the demographical section at the beginning of the survey. The familiarity item was displayed on a 5-point Likert scale ranging from 'not familiar at all' to 'extremely familiar' (Appendix B). Specifics on the Qualtrics survey can be found in Appendix C, which displays the complete survey.

2.2.2 Researcher script

In order to keep the study as replicable and transparent as feasible, the session was based on a script that the researcher had to follow. The researcher script can be found in Appendix C.

2.2.3 Google Hangouts

Google Hangouts was used to meet with the participants online. Participants who signed up via SONA were sent an email containing a link to join the Google Hangouts meeting on the scheduled date and time. Participants that were recruited elsewhere were sent the link to the meeting via another web based medium. Google Hangouts was used during the complete participation process and was used to record the sessions. The Qualtrics survey was shared with participants in the Google Hangouts meeting.

2.2.4 Chatbots and tasks

The Qualtrics survey flow was set up in a way that automatically assigns five out of the twelve chatbots to participants. In other words, every participant tested a different subset of five chatbots, and therefore only the relevant tasks and questions were presented. The twelve chatbots that were part of this study are ATO, HSBC UK, Absolut, Booking.com, USCIS, Emirates Holidays, Hubspot, Amtrak, Utwente, NBC News, ManyChat, and Job bot. Accompanying tasks per chatbot can be found in Appendix C.

2.3 Procedure

This study was conducted in an online environment due to safety considerations concerning the COVID-19 pandemic. As a first step, participants received a link to a Google Hangouts Meeting five minutes prior to their scheduled participation. When participants joined the meeting they were given a short introduction into the study and the meaning of their participation (Appendix C). Thereafter, the link to the Qualtrics survey was sent to participants (Appendix C). Before interacting with the chatbots, some pre-experimental data was obtained via the provided Qualtrics survey. For starters, participants filled in a informed consent form, stating their anonymous and confidential contribution to the study. In addition, they were provided with information about the purpose of the study and on the process of analyses on their data. After confirmation of participants' informed consent, additional demographical information was obtained from the participants, like their age, gender, study background, chatbot familiarity, and nationality. Once this pre-experimental data was obtained via Qualtrics, and there were no further questions, the study could begin.

Each chatbot was tested through two tasks, which participants had to try to complete in order to assess their satisfaction with a certain chatbot. After completing the first task, a question was posed regarding the level of ease that was experienced while performing this task. Thereafter, the second task was presented to the participant, and the same question was posed after completing the task. After completing both tasks, the USQ and the SUISQ-MR (i.e. 51 questions) were posed to assess satisfactory perceptions on the chatbot interaction. This process of filling in the questionnaires after interacting with a chatbot (i.e. completing the two stated tasks) was repeated five times per participant. In other words, every participant interacted with five chatbots, meaning that they had to conduct ten tasks and answer 255 questions in total. After the interaction with the chatbots, one final questionnaire (15 items) was posed that measured interest in technology/geekism. The responses of the participants were automatically saved in the Qualtrics platform during and after completion of the session.

In order to maintain transparency and replicability of the study, audio recordings and screen recordings (if possible) were taken. After the pre-experimental data were obtained, these audio- and screen recordings were initiated. Hence, all interactions of the participant, with either chatbot or researcher, were recorded. In the end, participants were thanked for their participation and were given the opportunity to ask questions.

2.4 Data analysis

R statistics was used to conduct the analyses that are relevant for this study. Before analyses took place the data had to be cleaned and checked for outliers. Microsoft Excel was used for cleaning the data and for the computation of total USQ-, SUISQ-MR-, and Geekism scores. Thereafter, further computations were performed in R statistics.

Explorative and descriptive statistics were used to gain insights regarding partcipants characteristics. Factor extraction of the USQ was conducted by means of a principal component analysis with an oblique (oblimin) rotation method. Bartlett's test of sphericity, and the Kaiser-Meyer-Olkin (KMO) measure were used to check the assumptions prior to the conduction of the principal component analysis. To be more specific, Bartlett's test of sphericity should be statistically significant, whereas the KMO value should be at least 0.5 in order to conduct the PCA (Kaiser, 1974). After checking these criteria, initial factor extraction was based on the examination of factors with an Eigenvalue of at least one. Because research has shown that such initial factor extraction may be erroneous (Ledesma & Valero-Mora, 2007), additional insights were gained via conducting a parallel analysis. In turn, these were compared to expectations set by previous research.

Also during the conduction of the PCA, some criteria were checked. Namely, during PCA, items were removed that had a commonality score below 0.2. Costello and Osborne (2005) illustrate that this cut-off criterion supports the removal of items that do not contribute significantly to their components. Later on, before structure matrices were established, an additional criterion regarding item loadings was added. In specific, item loadings of at least 0.3 were considered acceptable. As a result, item loadings below this criterion were removed. This criterion is in line with previous research that used similar criteria for establishing structure matrices (Dehmel & Borsci, 2020; Böcker & Borsci, 2019). Thereafter, additional cut-off criteria were manually checked. As such, items were manually removed that did not load 0.5 or higher on at least one component. Items were also removed if they cross-loaded highly with other components (i.e. loadings of 0.4 or higher on at least two other components). Hence, components consisting of items that loaded at least 0.5 on that component, without many high cross-loadings on other components, were considered valid principal components (Costello & Osborne, 2005).

Reliability analyses of the components that resulted from the PCA were performed by the computation of Cronbach's Alpha scores. These Cronbach's Alpha scores provided insight about the internal consistency of the USQ and the extracted components. Thereby,

Cronbach's Alpha scores below 0.65 were seen as insufficient (Dukes, 2005). The reliability analyses supported the removal of deficient items; items that decreased a component's internal consistency, in order to arrive at a revised and compressed version of the USQ.

The retained set of items, as well as individual established factors, were used to perform a correlation analysis with the SUISQ-MR to examine concurrent validity. More specifically, a significant correlation between scores on the SUISQ-MR and the USQ indicated sufficient (concurrent) validity. The Shapiro-Wilk test was performed to establish the relevant correlational analysis. The relevant correlational analysis, Kendall's Tau, was then conducted and supported by the calculation of 97.5% confidence intervals using a bootstrapping method with 9999 replicates.

Lastly, familiarity with chatbots and interest in technology (geekism) were analysed on a possible influence on USQ scores by means of non-linear local regression analyses. Non-linear local regression was used since the data did not meet linearity assumptions. Furthermore, these regression analyses have shown to be useful for both continuous and discrete numeric data, such as the familiarity- and geekism data (U.S. Environmental Protection Agency, 2016).

3. Results

3.1 Data cleaning

Data preparation was performed in Microsoft Excel. Before relevant total scores could be computed some negatively phrased items were reversed (indicated with an asterisk in Appendix B). After inversing the relevant items, total scores were computed per participant. To be more specific, every participant's individual total scores on the USQ, SUISQ-MR, and Geekism-questionnaire were established, where higher scores indicated more positive attitudes towards chatbots or technology in general. Responses on familiarity were recoded into numerical variables, in order to be able to compare familiarity scores with USQ scores.

The data exploration in R statistics led to the detection of two outliers. Both outliers were found in the item on familiarity with chatbots. In specific, only two out of 48 participants stated to be extremely familiar with chatbots. However, because these two responses seemed authentic, and therefore not due to an error or problem with the proper comprehension of the item, they were retained.

No missing values were detected. The likelihood of detecting missing values was already negligible due to the usage of a forced response option in Qualtrics. The only way that missing values could have been detected was if participants had prematurely ended their participation. If this was the case, these participants' contributions were already removed in Excel by checking participants' completion rate.

3.2 Principal Component Analysis

The following assumptions for a principal component analysis were tested. Firstly, it was checked whether the inter-item correlations were acceptable. In other words, a minimum of sizable correlations in the correlation matrix of the items of the USQ should have been observed. The correlation matrix showed that each item had at least one correlational value of ≥ 0.3. Therefore, all individual items met the first criterion of acceptable inter-item correlation. Secondly, the Kaiser-Meyer-Olkin indicated an overall KMO of 0.61, whereas the majority of individual items illustrated a sufficient MSA value (≥ 0.5). The established KMO value indicated that the sampling adequacy is mediocre but acceptable (Kaiser, 1974). As the last criterion, Bartlett's test of sphericity was conducted and was found to be significant with $x^2\ (861) = 2322.29$, $p < .001$. Hence, the three criteria were met, verifying conducting a PCA on the data.

In comparison to previous studies, it was expected that the initial factors to retain would lie in between four (Silderhuis & Borsci, 2020; Balaji & Borsci, 2019) and five

(Dehmel & Borsci, 2020; Böcker & Borsci, 2019). It was checked whether these expectations were met by means of parallel analysis. The parallel analysis illustrated that, based on Kaiser's criterion, the number of factors to retain was likely to be between four or five (Figure 2). Due to this insight, and the fact that communalities were higher for five factors than for four factors, a conservative option of five factors was opted for in further analyses.

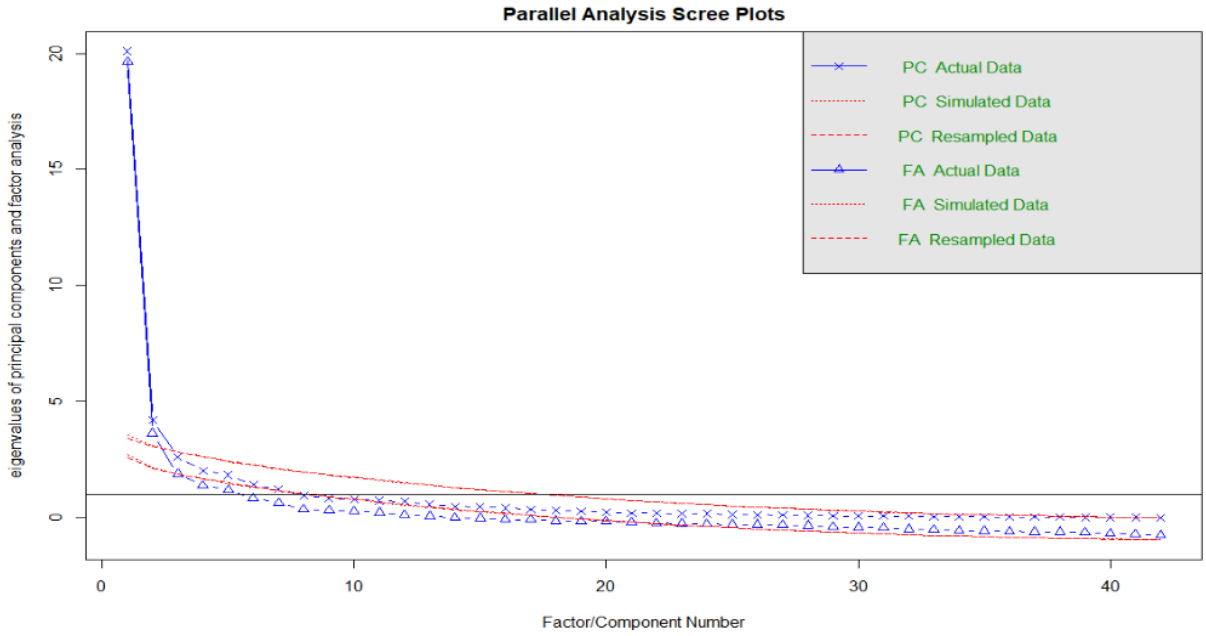


Figure 2. Scree plot of the parallel analysis on all 42 items

Thus, principal component analysis was first repeated for a fixed number of five components. The structure matrix that resulted from the oblique (oblimin) rotation allowed examination of item loadings per component. Factor loadings below 0.3 were automatically removed as data to-be-viewed in the structure matrix. Moreover, items were manually removed that did not load ≥ 0.5 on at least one component. Items were also removed if they cross-loaded highly with other components (i.e. loadings of 0.4 or higher on at least two other components). Hence, components consisting of items that loaded at least 0.5, without many high cross-loadings on other components, were retained. The obtained structure matrix can be found in Table 2, where retained items per component appear in bold.

Table 2.

Oblique factor loadings with five components

	TC1	TC2	TC3	TC4	TC5
USQ1	0.38	**0.88**		0.55	
USQ2	0.4	**0.91**		0.39	
USQ4	0.41	**0.88**		0.35	
USQ5	0.34	**0.79**		0.4	
USQ6	0.38	**0.79**	0.31		
USQ8	**0.7**	0.63			
USQ9	**0.68**	0.58			
USQ10	0.59				**0.73**
USQ11					**0.84**
USQ13	0.45		**0.56**		
USQ17		0.41		**0.66**	
USQ19	0.33		**0.82**		
USQ20			**0.89**		
USQ21	0.34		**0.84**		
USQ22	**0.89**	0.36	0.32	0.48	0.39
USQ23	**0.77**			0.53	
USQ24	**0.92**	0.33		0.41	0.31
USQ26	**0.89**		0.36		
USQ27	**0.92**	0.46	0.3	0.3	
USQ28	**0.81**		0.45		
USQ29	**0.88**	0.32	0.34		
USQ30	**0.84**	0.37	0.44	0.34	
USQ31	**0.55**				0.52
USQ37	**0.86**	0.54		0.35	
Eigenvalues	13.34	7.18	3.41	4.32	2.53
% Explained	43	23	11	14	8
Cronbachs' α	0.96	0.93	0.83	NA	0.84

The structure matrix illustrated that one factor (TC4) consisted of only one item. For this reason, it was not possible to gain a valid insight into this component's reliability. Since the parallel analysis indicated that a structure of the questionnaire composed of four factors, instead of five, could be acceptable, it was examined whether this alternative composition would induce better results.

For this analysis, the same criteria were adopted with regard to data to-be-viewed in the structure matrix (i.e. factor loadings above 0.3 are displayed, loadings above 0.5 are eligible items for representing components, and not more than one cross-loading should be above 0.4). Repeating this same analysis for a fixed number of four components, with an oblique rotation method, resulted in the following structure matrix (Table 3).

Table 3.

Oblique factor loadings with four components

	TC1	TC2	TC3	TC4
USQ1	0.32	**0.88**		0.55
USQ2	0.32	**0.91**		0.44
USQ3	0.38	**0.88**		0.42
USQ4	0.42	**0.88**		0.3
USQ5	0.34	**0.8**		0.35
USQ6	0.35	**0.79**		
USQ8	**0.69**	0.64		0.33
USQ9	**0.68**	0.6		0.3
USQ10	**0.77**			
USQ11	**0.51**			
USQ12	**0.85**	0.42		
USQ13	**0.51**		0.45	
USQ14	**0.78**	0.41	0.37	0.38
USQ17		0.41		**0.6**
USQ19	0.31		**0.71**	
USQ20			**0.84**	
USQ21	0.32		**0.75**	
USQ22	**0.92**	0.38		0.49
USQ23	**0.77**			0.56
USQ24	**0.93**	0.35	0.33	0.47
USQ26	**0.85**		**0.58**	0.37

USQ31	**0.66**			
USQ32		0.43		
USQ42	0.39	0.46		**0.8**
Eigenvalues	12.25	7.54	4.01	5.14
% Explained	42	26	14	18
Cronbachs' α	0.93	0.94	0.79	0.65

The obtained structure matrix on four principal components illustrated that no factors were composed of only one item. For that reason, opting for four factors appeared to be a verifiable choice. Further inspection of the obtained components showed that all components had a Cronbach's alpha above 0.65 and that there were no item-rest correlations below 0.3, which indicated that all 24 items in the structure matrix could be retained.

Hence, four components were established, each containing at least two items. Component 1 contains twelve items (i.e. 8, 9, 10, 11, 12, 13, 14, 22, 23, 24, 26, 31), component 2 contains six items (i.e. 1, 2, 3, 4, 5, 6), component 3 contains four items (i.e. 19, 20, 21, 32), and component 4 contains two items (i.e. 17 and 42). Component 1 includes items with features that represent communication quality, like 'recognition and facilitation of users' goal and intent', 'communication effort', 'ability to maintain themed discussion', and 'expectation setting'. Component 2 includes items with features that represent accessibility and ease of getting started, which are 'accessibility' and 'ease of starting a conversation'. Component 3 includes items with features that represent perceived privacy and graceful handling, which are 'perceived privacy' and 'graceful breakdown'. Component 4 includes items with features that represent response time and reference usage, which are 'reference to the relevant service' and 'perceived speed'.

Consequently, an overall insight in the established components, their content and accessory features (Balaji & Borsci, 2019), can be found in Table 4. Concurrently, this table also illustrates the final version of the USQ, with four factors and 24 items in total.

Table 4.

Factorial structure of the USQ

Factor	Item	Item content	Feature
1: Communication quality	USQ_8	I was immediately made aware of what information the chatbot can give me.	F3: Expectation setting

	USQ_9	It is clear to me early on about what the chatbot can do.	F3: Expectation setting
	USQ_10	I had to rephrase my input multiple times for the chatbot to be able to help me.	F4: Communication effort
	USQ_11	I had to pay special attention regarding my phrasing when communicating with the chatbot.	F4: Communication effort
	USQ_12	It was easy to tell the chatbot what I would like it to do.	F4: Communication effort
	USQ_13	The interaction with the chatbot felt like an ongoing conversation.	F5: Ability to maintain themed discussion
	USQ_14	The chatbot was able to keep track of context.	F5: Ability to maintain themed discussion
	USQ_22	I felt that my intentions were understood by the chatbot.	F8: Recognition and facilitation of users' goal and intent
	USQ_23	The chatbot was able to guide me to my goal.	F8: Recognition and facilitation of users' goal and intent
	USQ_24	I find that the chatbot understands what I want and helps me achieve my goal.	F8: Recognition and facilitation of users' goal and intent
	USQ_26	The chatbot is good at providing me with a helpful response at any point of the process.	F9: Relevance
	USQ_31	The chatbot could handle situations in which the line of conversation was not clear.	F11: Graceful breakdown
2: Accessibility and ease of getting started	USQ_1	It was clear how to start a conversation with the chatbot.	F1: Ease of starting a conversation
	USQ_2	It was easy for me to understand how to start the interaction with the chatbot.	F1: Ease of starting a conversation
	USQ_3	I find it easy to start a conversation with the chatbot.	F1: Ease of starting a conversation
	USQ_4	The chatbot was easy to access.	F2: Accessibility
	USQ_5	The chatbot function was easily detectable.	F2: Accessibility
	USQ_6	It was easy to find the chatbot.	F2: Accessibility
3: Perceived privacy and graceful handling	USQ_19	The interaction with the chatbot felt secure in terms of privacy.	F7: Perceived privacy

	USQ_20	I believe the chatbot informs me of any possible privacy issues.	F7: Perceived privacy
	USQ_21	I believe that this chatbot maintains my privacy.	F7: Perceived privacy
	USQ_32	The chatbot explained gracefully when it could not help me.	F11: Graceful breakdown
4: Response time and reference usage	USQ_17	The chatbot is using hyperlinks to guide me to my goal.	F6: Reference to the relevant service
	USQ_42	The chatbot is quick to respond.	F14: Perceived speed

After the final 24-item version of the USQ was proposed, the results were then compared with structures obtained from previous studies (Dehmel & Borsci, 2020; Silderhuis & Borsci, 2020; Balaji & Borsci, 2019; Böcker & Borsci, 2019). Table 5 provides a comparison among the 24-item USQ that emerged in the current study.

Table 5.

Comparing the 24-item USQ with previous findings

Dehmel and Borsci (2020)	Silderhuis and Borsci(2020)	Balaji and Borsci (2019)	Böcker and Borsci (2019)	24-version USQ
Quality and quantity of information: 16, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 37, 38, 39.	Communication quality: 7, 8, 9, 12, 13, 14, 15, 16, 18, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 34, 35, 37, 39.	Response quality: 7, 15, 18, 24, 25, 30, 33, 34, 37.	General usability: 8, 10, 11, 12, 14, 22, 23, 24, 26, 27, 29, 31, 37.	Communication quality: 8, **9**, **10**, **11**, 12, **13**, 14, 22, 23, 24, 26, 31.
Ease of getting started: 1, 2, 3, 4, 5, 6.	Conversation start: 1, 2, 3, 4, 5, 6	Communication quality: 1, 2, 4, 5, 10, 11.	Ease of getting started: 2, 3, 4, 5, 6.	Accessibility and ease of getting started: 1, 2, 3, 4, 5, 6.
Perceived privacy and security: 19, 20, 21.	Perceived privacy: 19, 21.	Perceived privacy: 21.	Perceived privacy and security: 19, 20, 21.	Perceived privacy and graceful handling: 19, 20, 21, **32**.
Response time: 40, 41, 42.	Perceived speed: 41, 42.	Perceived speed: 41.	Response time: 40, 41, 42.	Response time and reference usage: **17**, 42.
Keeping track of context: 13, 14, 31, 32, 33.			Articulateness: 33, 35, 36.	

Note. The numbers in bold represent the items that were different in at least three other studies.

3.3 Correlational analysis between the 24-item USQ and the SUISQ-MR

To inspect the relationship between the 24-item USQ and the SUISQ-MR, a Kendall's Tau correlational analysis was performed, as the Shapiro-Wilk test indicated that the data was not normally distributed. Results indicated that the total scores of the two questionnaires correlated with $r_t = 0.342$ and $p < 0.001$. Further bootstrapping with 9999 replicates indicated that there is a 97.5% confidence that the true correlation value lays in between [0.15 ; 0.51] (Figure 3).

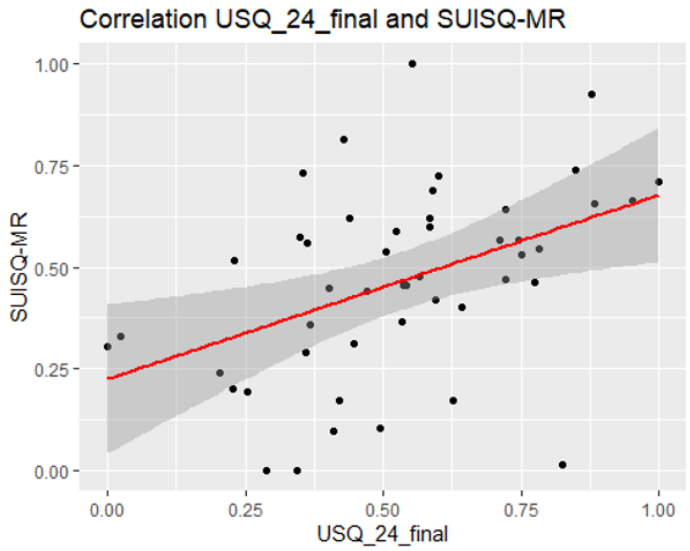


Figure 3. Correlation between final USQ and SUISQ-MR

Thereafter, Kendall's Tau test was performed to inspect relationships between individual factors of the 24-item USQ and the SUISQ-MR. Table 6 illustrates the significant correlations between the two questionnaires and their factors.

Table 6.

Significant correlations between final USQ factors and SUISQ-MR factors

	F1: User Goal Orientation (SUISQ-MR)	F2: Costumer Service Behaviors (SUISQ-MR)	F3: Speech characteristics (SUISQ-MR)	F4: Verbosity (SUISQ-MR)	Overall score (SUISQ-MR)
F1:Communication quality (USQ)	0.453***	0.251*	-	0.343***	0.305**
F2: Accessibility and ease of getting started	0.222*	0.309**	-	0.225*	0.322**

(USQ)					
F3: Perceived privacy and graceful handling (USQ)	0.240*	0.285**	-	-	-
F4: Response time and reference usage (USQ)	-	0.274**	-	-	-
Overall Score (USQ)	0.456***	0.333**	-	0.342***	0.342***

Note. *p<.05, **p<0.01, ***p<0.001

3.4 Non-linear regression analyses

Non-linear local regression analyses were conducted in order to examine whether there were significant relationships between chatbot familiarity and overall scores on the final USQ, and between geekism and overall scores on the final USQ. More specifically, the overall score on the item that measures familiarity and the overall score on the Geekism-questionnaire were checked for a (non-linear) relationship with the overall score on the final proposed version of the USQ.

The results indicated that familiarity was identified as a significant predictor of the overall score on the final USQ. In specific, familiarity had a weak upward-opening parabolical relationship with the overall score on the final USQ with a correlation estimate of 0.26 with a $p < 0.01$.

There was no significant non-linear relationship between geekism and the overall score on the final USQ. In specific, the explored relationship indicated that geekism was negligibly associated with the overall score on the final USQ, in a non-significant manner.

4. Discussion

The goal of this research was to arrive at a refined USQ that measures an overall experience with textual chatbots. In specific, the 24-item USQ-structure was established and compared to previously proposed structures, the 24-item USQ was examined on its reliability and validity, and possible confounding effects were studied. The following paragraphs will focus on interpreting the results via answering the research questions.

4.1 Factorial USQ-Structure

In this section the first research question: "Does the current study confirm previously proposed factorial USQ-structures?" will be answered. The PCA resulted in a USQ composed of 24-item with a four-factor structure. Comparing this result with previously established structures led to the observation that most of the retained items were also retained in previous research. Specifically, 83.3% of the items that were included in the USQ were also retained in at least two other previous studies. Therefore, it can be stated that the overall 24-item USQ confirms previously proposed factorial USQ-structures.

To gain more insight about the overlap between other proposed USQ-structures and the 24-item USQ, the individual factors were compared as well. A first note regards that there were two items (i.e. item 10 and item 11) that were merged differently in other studies. More specifically, in this study, item 10 and item 11 were combined in similar factors to the study by Böcker and Borsci (2019), whereas in the study by Balaji and Borsci (2019) these items were combined in different factors. Considering the content and feature representation of item 10 and item 11 (Appendix B) it can be hypothesized that item 10 and 11 have a somewhat better representation in the factor 'communication quality' in the current study, as opposed to the composition in which these items are represented by Balaji and Borsci (2019). Beside this observation, the factorial structure and the content of factors overlapped for 75% with at least two other studies. Therefore, it can be stated that the overall 24-item USQ, as well as its factors, overlaps significantly with previous research. This finding strengthens the findings of previous research as well as the findings of this research.

4.2 Reliability and concurrent validity of the final USQ

Reliability assessment of the final USQ was based on the calculation of Cronbach's Alpha scores for the overall final USQ as well as for the individual factors. The 24-item USQ had an overall Cronbach's Alpha score of 0.93 which indicates that the final proposed USQ consists of an excellent internal consistency (Hair, Babin, Money, & Samouel, 2003). The first and second factor, 'communication quality' and 'accessibility and ease of getting

started', both consist of excellent internal consistency (Hair et al., 2003). The third factor, 'perceived privacy and graceful handling', consists of a good internal consistency (Hair et al., 2003). The fourth factor, 'response time and reference usage', consists of a moderate internal consistency (Hair et al., 2003). Hence, the overall final USQ, as well as its factors, consist of sufficient reliability to be implemented as a reduced version of the original 42-item USQ.

Moving on to concurrent validity assessment, the following section will focus on answering the second research question: "Do the results of the final proposed USQ correlate with the results of the SUISQ-MR?" The answer to this research question is based on the results of the correlational analysis between the 24-item USQ and the SUISQ-MR and their individual factors (Table 6). Overall, the scores on the SUISQ-MR and on the 24-item USQ correlated significantly. In specific, a positive relationship between the scores of the two questionnaires was identified. This finding supports the establishment of sufficient concurrent validity and therefore supports the implementation of this questionnaire as a measure of user satisfaction with chatbots. In previous research, another questionnaire, the UMUX-Lite (Lewis, Utesch, & Maher, 2013), was utilized to assess the concurrent validity of the USQ (Dehmel & Borsci, 2020; Silderhuis & Borsci, 2020; Balaji & Borsci, 2019; Böcker & Borsci, 2019). Hence, the utilization of the SUISQ-MR, as opposed to the UMUX-Lite, strengthens previous establishments of concurrent validity of the USQ.

4.3 Confounding influences of chatbot familiarity and geekism

This section will provide an answer to the third and fourth research question: "Does familiarity with chatbots affect participants' answers to the USQ?" and "Does interest in technology (geekism) affect participants' answers to the USQ?". The answers to these research questions are based on non-linear regression analyses. Similar to Dehmel and Borsci (2020), this study indicated that geekism had no significant influence on the overall USQ score. Since the 24-item USQ differs from the revised USQ by Dehmel and Borsci (2020), this finding strengthens the belief that geekism does not influence overall satisfaction with chatbots.

In contrast to Dehmel and Borsci (2020), in the present study, a significant influence of chatbot familiarity on overall USQ scores was identified. Although this was a weak relationship it might be an important confounding variable that contributes to significant differences in user satisfaction with chatbots. In specific, the upward-opening parabolic relationship indicates that overall user satisfaction with chatbots was lowest for participants with moderate familiarity with chatbots and higher for participants that indicated to be below

or above moderately familiar with chatbots. This finding is partly in line with previous researchers that stated that a higher familiarity with technology positively influences perceived usability and performance (Fu & Gray, 2004; Tuch et al., 2012; Payne et al., 2000). Additionally to the affirmation of this positive relationship, the finding that non-familiar and slightly familiar participants rated their satisfaction higher as opposed to moderately familiar participants, was unexpected. A possible explanation for this finding might lay in the fact that some participants without familiarity with chatbots mentioned that they expected the tasks to be harder to achieve and were surprised by certain chatbot abilities. As this possible explanation was not further studied, it can only be hypothesized that this may be an explanation of why non-familiar participants were more satisfied with their experiences than moderately familiar participants.

4.4 Limitations and future recommendations

4.4.1 Sample limitations

The BMS Test Subject Pool system (SONA) and convenience sampling were used in this study to recruit participants. These sampling methods were beneficial in terms of study costs and time consumption. However, these sampling methods may bring weaknesses to the generalization of study results. In specific, this study might be susceptible to selection bias, as not all people had an equal chance to participate in this study (Taherdoost, 2016).

Further limitations can be detected in the descriptive statistics of the sample. The mean age of participants was 21.9 years old, which is relatively low when aiming for generalization to a wider age range. This observation can be explained by the fact that the majority of participants were students who were recruited via SONA.

4.4.2 COVID-19 limitations

Since this study was conducted during the COVID-19 pandemic, some limitations should be noted concerning participants' interaction with the chatbots. Due to COVID-19 and accompanying safety considerations, this study was conducted online. Therefore, a first possible limitation is that participants had to interact with chatbots via their own computer devices and in their own personal environments. For that reason, perceived experiences with chatbots might have been influenced by the personal circumstances and resources of the participant, such as computer quality, environmental distractions, and internet connection.

4.4.3 Chatbot limitations

In this study, a total number of twelve chatbots was used. In theory, only a set of five chatbots could have been sufficient since each individual participant assesses only five out of twelve chatbots. Yet, using a set of twelve chatbots instead of five increases the likelihood that the 24-item USQ measures user satisfaction for all chatbots, instead of solely the set used in this study. However, although the usage of a set of twelve chatbots supports generalization to other chatbots, some limitations should be noted.

It should be highlighted that participants might have had previous experiences with some of the chatbots. As an example, many participants noted that they were familiar with the Booking.com website. Although not many of those participants interacted with the Booking.com chatbot before, they did mention to have a sense of how to gain the information they needed.

Four chatbots operated via Facebook, which may indicate that participants interacted intuitively with those chatbots since the majority of participants was familiar with Facebook Messenger. Therewith, it should also be noted that if participants were assigned to multiple chatbots that operate via Facebook, a learning effect might have occurred that influenced their perception. Hence, participants' perceptions of their user satisfaction with a second chatbot that operates via Facebook, might not be valid.

Lastly, it should be noted that participants assessed different subsets of chatbots. As discussed in paragraph 2.2, participants were randomly assigned to five out of twelve chatbots. For that reason, every participant interacted with a different subset of five chatbots. Besides the benefits that were accompanied by this choice, this may also induce some limitations on comparisons between participants. As an example, some chatbots were experienced to be harder to interact with. Therefore, some secondary effects in results might have occurred between participants who were assigned to these chatbots and participants who were assigned to, for example, multiple Facebook-operating chatbots.

4.4.4 24-item USQ limitations

During participants' interaction with the chatbots, some item-related concerns became apparent. Namely, items 5, 17, 20, 21, 24, and 32 often required clarification for participants. With regard to item 5 it was often asked what was meant by 'chatbot function'. Items 17, 24, and 32 were accompanied by issues due to their non-exhaustive nature. In specific, for item 17 participants stated that they did not know what to answer if the chatbot used hyperlinks, but did not help them to achieve their goal. Items 24 and 32 illustrated the same ambiguity.

Participants did not know what to answer on item 24 if they felt that the chatbot understood what they wanted, but did not help with achieving their goal. With regard to item 32, participants did not know what to answer if the chatbot did provide them with an explanation of why it could not help them, yet without giving a graceful explanation. Thereby, also some concerns were stated for the exact meaning and interpretation of a graceful response. More concerns came to the fore regarding items 20 and 21, which measure participants' perceived level of privacy. Participants stated their concerns once they had to answer these questions for chatbots that operate via Facebook. Moreover, participants often stated to have little trust in any technology that operates via Facebook, and therefore did not really assess the chatbot's quality.

One last concern should be noted with regard to factor 4; 'response time and reference usage'. This factor consisted of two items, which might decrease its factor stability (Floyd & Widaman, 1995). Floyd and Widaman (1995) state that any factor containing less than three items, might be susceptible to a weak factor stability. In addition, since item 17 was perceived as somewhat ambiguous by participants, this could imply an additional issue for the quality of the fourth factor.

4.4.5 Future recommendations

Recommendations should be posed that aim to resolve the previously discussed limitations for future research. First, changes could be made in the composition and recruiting of participants. It is recommended that future research includes more people of different age categories, in order to arrive at a broader age-spectrum in the sample. Next to that, also more participants should be recruited that are at least very familiar with chatbots. The current study illustrated that there were only two participants who stated to be extremely familiar with chatbots, and were therefore treated as outliers.

Next, it is recommended to add a question in the demographical section of the Qualtrics survey regarding participants' pre-existing attitude towards chatbots. This question should only be asked to the participants that stated to be at least moderately familiar with chatbots. If future research consists of more chatbot-experienced participants, a valuable distinction could be made between participants who are familiar and enjoy chatbots and participants who are familiar without enjoying the interaction with chatbots. Thereby, it is also recommended to pose the Geekism-questionnaire at the beginning of the research. Participants might be influenced by their experiences with the chatbots if they fill out the questionnaire after the study is completed. Although this research indicated that geekism did

not influence overall user satisfaction with chatbots, this alternative explanation could be examined in future research.

Considering the subset of chatbots, it is recommended that future research only retains one of the chatbots that operates via Facebook. This recommendation aims to diminish a possible learning effect on how to interact with chatbots via Facebook (Messenger). Preferably, at least the Booking.com chatbot could be removed since familiarity with its website was discussed on a possible confounding effect on satisfaction.

Lastly, two recommendations are made that aim to improve the overall study quality. A first possible indication is to replicate this study without, or by correcting, the items that illustrated concerns by participants (i.e. without items 5, 17, 20, 21, 24, and 32). A second and last indication is to aim for a more standardized study environment. It is recommended to future researchers to assess how the study environment could be as similar as possible considering the COVID-19 pandemic.

5. Conclusion

The following section will focus on evaluating the main research goal; to arrive at a reliable, validated, and refined USQ that measures user satisfaction with textual chatbots. This study led to a proposal of a 24-item version of the USQ which consists of an overall excellent reliability score and sufficient (concurrent) validity. The 24-item USQ is useful for assessing an overall experience with chatbots, which was previously not possible. Furthermore, this study provided insight about usability concepts that are important for positive experiences with chatbots. Hence, the 24-item USQ fills the need for a standardized questionnaire to assess experiences with chatbots.

The majority of conclusions drawn in this study confirm previously established USQ-structures and therefore strengthen the results of previous research, as well as results of the current research. However, some different and unexpected results significantly add to the findings of previous research and which are promising to explore in future research, such as the influence of chatbot-familiarity on USQ scores.

Concludingly, although further exploration of an optimal revised version of the USQ is advised, the current study, combined with previous studies (i.e. Dehmel & Borsci, 2020; Silderhuis & Borsci, 2020; Balaji & Borsci, 2019; Böcker & Borsci, 2019), reveals that deployment of the USQ as a measure for user satisfaction with chatbots, is advancing.

Literature

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

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, *4*(2), 115-130.

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

Brandtzaeg, P. B., & Følstad, A. (2017, November). Why people use chatbots. In *International Conference on Internet Science* (pp. 377-392). Springer, Cham.

Coperich, K., Cudney, E., & Nembhard, H. (2017). Continuous improvement study of chatbot technologies using a human factors methodology. In *Proceedings of the 2017 Industrial and Systems Engineering Conference*.

Deci, E. L., & Ryan, R. M. (1985). Cognitive evaluation theory. In *Intrinsic motivation and self-determination in human behavior* (pp. 43-85). Springer, Boston, MA.

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

Duijst, D. (2017). Can we improve the User Experience of Chatbots with Personalisation. *Master's thesis. University of Amsterdam*.

Dukes, K. A. (2005). Cronbach's Alpha. *Encyclopedia of biostatistics, 2.*

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, *7*(3), 286.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *interactions*, 24(4), 38-42.

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

Fu, W. T., & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive science*, *28*(6), 901-935.

Goossens, F. (n.d.). *Designing a VUI - Voice User Interface*. Toptal Design Blog. https://www.toptal.com/designers/ui/designing-a-vui

Hair, J. F. Jr., Babin, B., Money, A. H., & Samouel, P. (2003). Essential of business research methods. John Wiley & Sons: United States of America.

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.

Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, *6*(3-4), 287-303.

International Organization for Standardization. (2010). *ISO DIS 9241-210:2010(en) Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.* ISO.

International Organization for Standardization. (2018). *ISO 9241-11:2018(en) Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. ISO. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:en

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British journal of educational technology*, *24*(3), 210-212.

Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016, July). Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 45-54).

Koopman, R., & Schmettow, M. (2019). *The Uncanny Valley as a universal experience: a replication study using multilevel modelling* (Bachelor's thesis, University of Twente).

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research, and evaluation*, *12*(1), 2.

Lewis, J., & Sauro, J. (2020, March 3). *Three questionnaires for measuring voice interaction experiences.* MeasuringU. https://measuringu.com/voice-interaction/

Lewis, J. R. (2017). Investigating MOS-X Ratings of Synthetic and Human Voices.

Lewis, J. R., & Hardzinski, M. L. (2015). Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology*, *18*(3), 479-487.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). 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). ACM

Macaranas, A., Antle, A. N., & Riecke, B. E. (2015). What is intuitive interaction? balancing users' performance and satisfaction with natural user interfaces. *Interacting with Computers*, *27*(3), 357-370.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22-32.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10, 1-9. doi:10.4135/9781412995627.d8

Payne, S. J., Richardson, J., & Howes, A. (2000). Strategic use of familiarity in display-based problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1685.

Piccolo, L. S., Mensio, M., & Alani, H. (2018, October). Chasing the chatbots. In *International Conference on Internet Science* (pp. 157-169). Springer, Cham.

Schmettow, M., & Drees, M. (2014, September). What drives the geeks? Linking computer enthusiasm to achievement goals. In *Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI 2014) 28* (pp. 234-239).

Shawar, B. A., & Atwell, E. (2007, January). Chatbots: are they really useful?. In *Ldv forum* (Vol. 22, No. 1, pp. 29-49).

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

Tariverdiyeva, G., & Borsci, S. (2019). *Chatbots' Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis* (Master's thesis, University of Twente).

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research (April 10, 2016)*.

Tuch, A. N., Presslaber, E. E., Stöcklin, M., Opwis, K., & Bargas-Avila, J. A. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International journal of human-computer studies*, *70*(11), 794-811.

U.S. Environmental Protection Agency. (2016, July). *Loess (or Lowess)*. https://www.epa.gov/sites/production/files/2016-07/documents/loess-lowess.pdf

Van Hooij, E. R. (2016). *Image schemas and intuition: the sweet spot for design?* (Master's thesis, University of Twente).

Appendices

Appendix A.

Table A1.

Key features on user perception with chatbots (Tariverdiyeva & Borsci, 2019)

	Key feature	Description
1	Response time	Ability of the chatbot to respond timely to users' requests
2	Maxim of quantity	Ability of the chatbots to respond in an informative way without adding too much information
3	Maxim of quality	Ability of the chatbot to avoid false statements/information
4	Maxim of manners	Ability of the chatbot to make its purpose clear and communicate without ambiguity
5	Maxim of relation	Ability of the chatbot to provide the relevant and appropriate contribution to people needs at each stage
6	Appropriate degrees of formality	Ability of the chatbot to use appropriate language style for the context
7	Reference to what is on the screen	Ability of the chatbot to use the environment it is embedded in to guide the user towards its goal
8	Integration with the website	Position on the website and visibility of the chatbot (all pages/specific pages, floating window/pull-out tab/embedded etc.)
9	Process facilitation and follow up	Ability of the chatbot to inform and update users about the status of their task in progress
10	Graceful responses in unexpected situations	Ability of the chatbots to gracefully handle unexpected input, communication mismatch and broken line of conversation
11	Recognition and facilitation of users' goal and intent	Ability of the chatbot to recognize user's intent and guide the user to its goal
12	Perceived ease of use	The degree to which a person believes that interacting with a chatbot would be free of effort
13	Engage in on-the-fly problem solving	Ability of the chatbot to solve problems instantly on the spot
14	Themed discussion	Ability of the chatbot to maintain a conversational theme once introduced and to keep track of the context to understand the user's utterances
15	Users' privacy and ethical decision making	Ability of the chatbot to protect user's privacy and make ethically appropriate decisions on behalf of the user

16	Meets neurodiversity needs	Ability of the chatbot to meet needs of users independently from their health conditions, well-being, age, etc.
17	Trustworthiness	Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage
18	Flexibility of linguistic input	Ability of the chatbot to understand users' input regardless of the phrasing

Appendix B.

B1. Familiarity item

5-point Liker scale of the familiarity item

1: Extremely familiar
2: Very familiar
3: Moderately familiar
4: Slightly familiar
5: Not familiar at all

Familiarity*: How familiar are you with chatbots and/or other conversational interfaces?

B2. Preliminary USQ

5-point Likert scale of the preliminary USQ

1: Strongly disagree

2: Somewhat disagree

3: Neither agree nor disagree

4: Somewhat agree

5: Strongly agree

USQ (key feature is in parentheses)

1. It was clear how to start a conversation with the chatbot. **(Ease of starting a conversation)**

2. It was easy for me to understand how to start the interaction with the chatbot. **(Ease of starting a conversation)**

3. I find it easy to start a conversation with the chatbot. **(Ease of starting a conversation)**

4. The chatbot was easy to access. **(Accessibility)**

5. The chatbot function was easily detectable. **(Accessibility)**

6. It was easy to find the chatbot. **(Accessibility)**

7. Communicating with the chatbot was clear. **(Expectation setting)**

8. I was immediately made aware of what information the chatbot can give me. **(Expectation setting)**

9. It is clear to me early on about what the chatbot can do. **(Expectation setting)**

10*. I had to rephrase my input multiple times for the chatbot to be able to help me. **(Communication effort)**

11*. I had to pay special attention regarding my phrasing when communicating with the chatbot. **(Communication effort)**

12. It was easy to tell the chatbot what I would like it to do. **(Communication effort)**

13. The interaction with the chatbot felt like an ongoing conversation. **(Ability to maintain themed discussion)**

14. The chatbot was able to keep track of context. **(Ability to maintain themed discussion)**

15. The chatbot maintained relevant conversation. **(Ability to maintain themed discussion)**

16. The chatbot guided me to the relevant service. **(Reference to service)**

17. The chatbot is using hyperlinks to guide me to my goal. **(Reference to service)**

18. The chatbot was able to make references to the website or service when appropriate. **(Reference to service)**

19. The interaction with the chatbot felt secure in terms of privacy. **(Perceived privacy)**

20. I believe the chatbot informs me of any possible privacy issues. **(Perceived privacy)**

21. I believe that this chatbot maintains my privacy. **(Perceived privacy)**

22. I felt that my intentions were understood by the chatbot. **(Recognition and facilitation of user's goals and intent)**

23. The chatbot was able to guide me to my goal. **(Recognition and facilitation of user's goals and intent)**

24. I find that the chatbot understands what I want and helps me to achieve my goal. **(Recognition and facilitation of user's goals and intent)**

25. The chatbot gave relevant information during the whole conversation. **(Relevance)**

26. The chatbot is good at providing me with a helpful response at any point of the process. **(Relevance)**

27. The chatbot provided relevant information as and when I needed it. **(Relevance)**

28. The amount of received information was neither too much nor too less. **(Maxim of quantity)**

29. The chatbot gives me the appropriate amount of information. **(Maxim of quantity)**

30. The chatbot only gives me the information I need. **(Maxim of quantity)**

31. The chatbot could handle situations in which the line of conversation was not clear. **(Graceful breakdown)**

32. The chatbot explained gracefully when it could not help me. **(Graceful breakdown)**

33. When the chatbot encountered a problem, it responded appropriately. **(Graceful breakdown)**

34. I found the chatbot's responses clear. **(Understandability)**

35. The chatbot only states understandable answers. **(Understandability)**

36. The chatbot's responses were easy to understand. **(Understandability)**

37. I feel like the chatbot's responses were accurate. **(Perceived credibility)**

38. I believe that the chatbot only states reliable information. **(Perceived credibility)**

39. It appeared that the chatbot provided accurate and reliable information. **(Perceived credibility)**

40. The time of the response was reasonable. **(Perceived speed)**

41. My waiting time for a response from the chatbot was short. **(Perceived speed)**

42. The chatbot is quick to respond. **(Perceived speed)**

B3. SUISQ-MR

7-point Likert scale ranging from strongly disagree (1) to strongly agree (7)

SUISQ-MR

1. I would be likely to use this system again.
2. I felt confident using this system.
3. The system used everyday words.
4. The system seems polite.
5. The system's voice sounded natural.
6. The system's voice sounded enthusiastic or full of energy.
7*. I felt like I had to wait too long for the system to stop talking so I could respond.
8*. The messages were repetitive.
9*. The system was too talkative.

B4. Geekism questionnaire

5-point Likert scale of the Geekism questionnaire

1: I totally disagree

2: I disagree

3: Cannot answer

4: I agree

5: I totally agree

Geekism questionnaire

1. I want to understand how computer parts and software work.
2*. Complex procedures with technical devices put me off.
3. I have sometimes modified a technical device or diverted it from its intended purpose.
4. I am motivated to optimize technical devices or configure them to my requirements.
5. I have, or I would make a project or work of mine publicly available on the internet.
6. Some people would call me a computer freak.
7*. I am not interested in the inner working or coding of software.
8. Challenging tasks with technical devices appeal to me.
9. I have good knowledge of computing devices.
10. I invest a lot of time and effort to explore computing devices.
11. I like acquiring more knowledge of technical devices.

12. I have more than once opened technical devices to see their insides.

13. Sometimes I use technical devices different to what they were intended for.

14*. It puts me off when technical devices have too many settings options.

15*. Usually, I need help when having trouble with a technical device.

Appendix C.

C1. Informed Consent Form

Start of Block: informed consent

Q96 **Informed consent**

Student investigator: Simone Wilmer

Q98 Thank you for participating in this study. This research is part of my Master thesis in Human Factors and Engineering Psychology at the University of Twente. The purpose of the research is to test and validate a preliminary questionnaire that measures user satisfaction with chatbot interaction. For this, you will interact with five chatbots and will perform two tasks for each. After that, the questionnaire as well as a measurement on task difficulty will be presented. Also some of your demographics will be filled out.

Demographic data will be used to see if certain characteristics, like previous experience with chatbots, have a significant effect on the experienced user satisfaction regarding the chatbots. The test will take between 30 minutes up to approximately one hour. Your test data will be processed anonymously. I do not anticipate that there are any risks associated with your participation, but you do have the right to stop the interview or withdraw from the research at any given time. The research project has been reviewed and approved by the BMS Ethics Committee.

The BMS Ethics Committee requires that academic research should meet ethical guidelines. These entail that participants explicitly agree on being interviewed, and that they agree on their participation as part of information gathering for academic purposes. This consent is necessary to ensure that you understand the purpose of your involvement and that you agree to the conditions of your participation. Therefore, I would like to suggest you read the accompanying informed consent form and answer the following statements. Please take note, if you agree to all of the following statements, you will continue to the questionnaire. However, if you disagree with any of the following statements, you will not partake in the study.

Q101 I consent voluntarily to be a participant in this study and understand that I can refuse to complete a task and I can withdraw from the study at any time, without having to give a reason.

- ○ Yes (1)
- ○ No (2)

Q103 I understand that taking part in the study involves an audio-recording as well as video recording, if possible. The data will be treated with discretion.

- ○ Yes (1)
- ○ No (2)

Q104 I agree to the collection of my age, gender, nationality, educational background and experience with chatbots. These data will be anonymized after this session.

- ○ Yes (1)
- ○ No (2)

Q105 I understand that the information I provide will be used for the Master thesis of the lead researcher of this study.

- ○ Yes (1)
- ○ No (2)

Q106 I understand that personal information collected about me that can identify me, such as [e.g. my age or where I live], will not be shared beyond the study team.

- ○ Yes (1)
- ○ No (2)

Q107 I have obtained sufficient information and was informed by the lead researcher in an appropriate manner. I understand that at the end of this survey, I will be given sufficient contact information in case of further questions.

- ○ Yes (1)
- ○ No (2)

Q108 Thank you for filling out the informed consent. The study will start now.

End of Block: informed consent

Start of Block: Introduction

C2. Researcher Script

Script

<<For researcher only: enter participant code>>

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

Please focus on achieving the tasks. At the end of each interaction, some questions will be asked on your experience with the chatbot. The session is expected to last for approximately one hour. I would like to record your voice and the screen for data analysis purposes. If you are not okay with this, please let us know. There are more details in the informed consent which you must read and sign before we move onto the actual chatbots. This informed consent is integrated in the online survey I will send you now.

<<Give participant informed consent form>>

First, please fill in the informed consent and demographic questionnaire.

You will now begin testing chatbots. Each provided task is a short realistic scenario – you, as the participant, should try your best to imagine yourself in those situations i.e. imagine that you're looking

for that information for the first time. If you do not understand the situation or task, let me know. **Once you feel like you have achieved the task, or if you feel that the task is not achievable, please let me know.**

Remember that I aim to assess the quality of the chatbot not you, if you cannot do something it's not your fault, but there is a problem with the tool. Also remember that there is no wrong or right answer in this experiment, I am mainly interested in what you think about the chatbot. Lastly, these chatbots may differ in quality so in some cases a human may step in, whereas in others this is less likely to happen.

Your behaviour and responses will help me understand how other users will experience these chatbots.

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

<<Start recording the screen>>

<<First chatbot: no new information>>

<<Chatbot 2 and 3 are tested>>

<<Chatbot 4 and 5 are tested>>

C3. Chatbots and tasks

Botname	Task 1	Task 2	Link	Defined state of success
Chatbot: ATO	You moved to Australia from the Netherlands recently. You want to know when the deadline is to lodge/submit your tax return using ATO's chatbot to find out.	You are a student and are wondering whether you have to lodge a tax return using the ATO's chatbot.	http://www.ato.gov.au/	Income tax Key lodgment and payment dates for business – income tax returns. These dates apply to entities that balance on 30 June, (at the end of the Australian financial year). They do not apply to entities that use a substituted accounting period. ---------- https://www.ato.gov.au/calculators-and-tools/do-i-need-to-lodge-a-tax-return/
HSBC UK	You live in the Netherlands but are travelling to Turkey for 2 weeks. During your travel, you would like to be able to use your HSBC credit card overseas at payment terminals and ATMs. You want to use HSBC's chatbot to find out the relevant procedure.	You have recently moved from Amsterdam to London and would like to know how you can change your address for your HSBC card, using the chatbot of HSBC UK.	https://www.hsbc.co.uk/	1. Notifying us is easy, all you need to do is: ---------- 1. Log into Online Banking 2. Hover over 'My Banking' 3. Select 'Notify us of Travel' 4. Select 'Create new travel plan' 5. Enter the details required and select 'Continue' 6. Review your input and if you're happy select 'Confirm' ----------

				2. Log on to Online Banking and select 'Personal & address details' within the 'My banking' menu at the top of the page. Your personal and address details - including your home and correspondence address - will immediately be displayed, and to update either of these, select 'Edit details' and follow the on-screen instructions. Please see our Change of Address support page for further information.
Absolut	You want to buy a bottle of Absolut vodka to share with your friends for the evening. One of your friends cannot consume gluten. You want to use Absolut's chatbot to find out if Absolut Lime contains gluten or not.	You want to buy a bottle of Absolut vodka for a good friend. But this friend is right now on a diet and tries to avoid sugar. You therefore want to find information about the amount of sugar in the products of Absolut using Absolut's chatbot.	https://www.absolut.com/en/	1. The gluten is removed in the production process since the spirit is first fermented and then distilled hundreds of times… ---------- 2. Absolut Vodka is Sugar free! Absolut Vodka does not contain carbohydrates, proteins or fat. This information is also valid for the flavoured products in our product range. For the production of flavoured vodka, only natural ingredients from berries, fruits and spices are used and no sugar is added.
Booking.com	You are travelling to London from 5th July to 9th July with your family. You want to use booking.com's chatbot to find a hotel room for you, your significant other and your child in Central London that does not cost more than 500€ in total	You have to attend an important business meeting from 18th to 19th of March in Amsterdam. You therefore are looking for a place to stay in the city center of London for not more than 200€ using the booking.com chatbot.	https://www.facebook.com/messages/t/1318400301782250	You got it. I found x properties for less than x Euros. Here are our top picks.
USCIS	You are a US citizen living abroad and want to vote in the upcoming federal elections. You want to use the USCIS chatbot to find out how.	You are planning to take a job in the USA. Since you are not a US citizen, you want to find out more about eligibility for a US- Green Card with the help of the USCIS	http://www.uscis.gov/emma	1. After you take the Oath of Allegiance at the citizenship ceremony, you will have the opportunity to register to vote. Voting registration forms may be distributed at your naturalization ceremony, After you become a U.S.

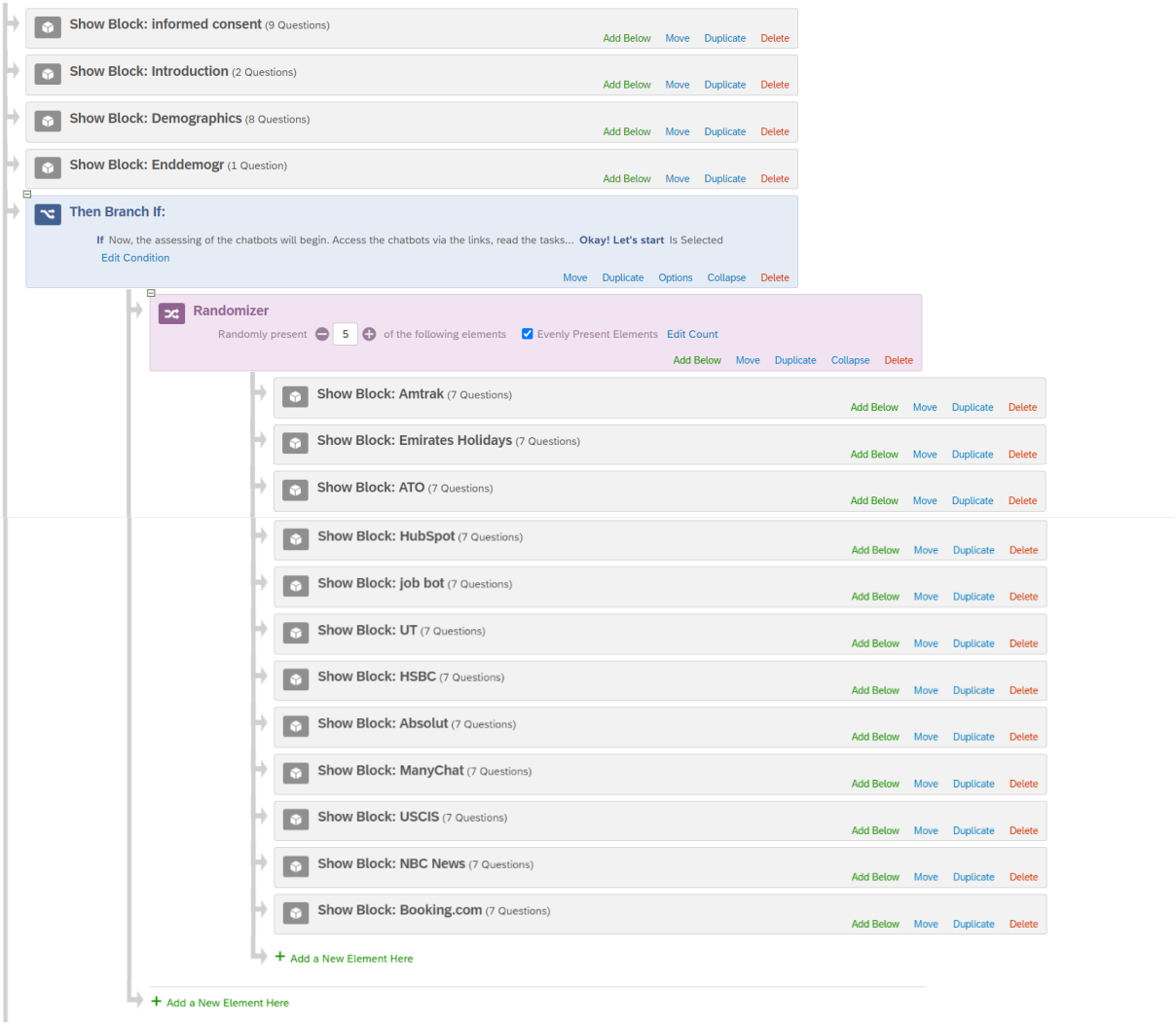
		chatbot..		citizen, you may also register to vote at other locations in your community, including post offices and motor vehicle offices. For more information, please see the links below. --- 2. In general, to meet the requirements for permanent residence in the United States, you must: Be eligible for one of the immigrant categories established in the Immigration and Nationality Act (INA); ...
Emirates Holidays	You just woke up and realize that you forgot that it's your significant other's birthday. Desperately, you are thinking about a birthday present and your idea is a holiday together in Paris. You visit the Emirates Holidays page and use Emirates Holidays' chatbot to book a holiday from the 4th September until the 9th September to Paris for two persons. Your departure airport is London Heathrow (LHR). Everything else is not important, as you just need a present for today.	You arrived in Paris and there seems to be a problem with your hotel reservation. You try to call someone at Emirates Holiday, but it's 11pm on Friday, so you cannot reach anyone. Hence, you ask Emirates Holidays' chatbot when the customer service opens on Saturday.	https://www.emiratesholidays.com/gb_en/	1) Chatbot asks for personal data (not needed to be given) and user indicated LHR as departure airport, Paris as destination, 2 persons are travelling, correct date and booking should be today --- 2) Hint: Click on x to end the current chat. Link to Opening hours page. Scroll down to customer service
Hubspot	You have your own company and would like to grow your business even more. A former colleague recommends you Hubspot. However, you don't want to sign up for anything (even if it's free). You use	Now, you are convinced that Hubspot can help your own business. Your focus is on improving your own customer service. Before you sign up for something, you would like to	https://www.hubspot.com/?survey=123	1) Hubspot Blog (after clicking on educational content) --- 2) Learn about products - provide better service - read more - link to page with heading 'Bring Order to Customer Service Chaos'

	Hubspot's chatbot to purely get information and get educated without using any tools. A collection of news/articles/tips would be great.	know how Hubspot can improve your customer service. You use Hubspot's chatbot to get more information about this.		
Amtrak	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.	You have planned a trip to the USA. You are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet an old friend for a few hours and see the city. You want to use Amtrak's chatbot to find out how much it will cost to temporarily store your luggage at the station.	https://www.amtrak.com/home	1) BBY and One-Way chosen with correct date. Shortest trip not possible with chatbot, only with sort function (make participant aware of chatbot failure in that sense) ---------- 2)' At-Station Baggage Services' page
Utwente	You are a chinese student who would like to do a Master's degree at the University of Twente. Your name is Jackie/Lin and your Email address is abc@def.com. You are interested in doing your master in Nanotechnology in September 2021. You did your bachelor at the Utwente in the Netherlands. You ask the Utwente chatbot what options for a	You are a german student who would like to do a Master's degree at the University of Twente. Your name is Alan/Sabine and your Email address is abc@def.com. You are interested in doing your master in computer science in February 2022. You did your bachelor's at the Jacobs University in	https://www.utwente.nl/en/education/master/chat/	1) Type in the information and follow the chatbot. Say yes at transfer minor or completed course and at question about scholarship. Click on arrow ---------- 2) Type in the information and follow the chatbot. Say yes at question about admission process and click on arrow.

	scholarship are available.	Bremen. You ask the Utwente chatbot about deadlines and the admission process.		
NBC News	You want to use the chatbot of NBC News to find out the most recent news regarding the environment.	Just out of curiosity, you are also interested in the most recent special coverage, using the chatbot of NBC News.	https://www.facebook.com/NBCNews/	Here's the latest from "Environment" ---------- Here is our top ongoing special coverage:
ManyChat	You want to integrate a chatbot on your companies' website. Therefore, you want to use the ManyChat's chatbot to find video tutorials to learn the basics of ManyChat.	After using the Chatbot for a while, you are getting a little bored and want to have some fun. Let the ManyChat's chatbot tell a joke to you.	https://www.messenger.com/t/ManyChat	Play tutorial #1 ---------- I've been programmed to share randomized jokes on my favorite topics. 🐙🤖💻 Be warned... I'm about to bring the cheese! 🧀
Job bot	You are looking for a new job as a teacher. Therefore, you want to use the Job bot to show you recent job offers in Manchester. You also want the results to be sorted by date	You want to see if the company behind the chatbot is a serious company. Therefore, you want to ask the chatbot, who developed him	https://www.messenger.com/t/jobbot.me	1. Start search, location: Amsterdam, teacher, offers will be shown, then click on sort by date 2. Type info. Asking the bot things like "who is your developer" or "who created you" or "where are you from" will make him answer things like: The internet, my developer etc.
Potential extra bots:	Facebook: Molly Mahoney - The prepared performer (not that great, but better than nothing) Kaimana Jerky (extremely limited) **Job Bot (pretty good bot)**	Others: Tidio (kind of a combination of HubSpot and ManyChat; https://www.tidio.com/) Mitsuku (conversational AI; no real purpose; might not be comparable to other chatbots, but is a good bot; https://pandorabots.		

		com/mitsuku/) Cleverbot (good, if we want a chatbot that talks about random things and insults the participant :)) **Heek (interesting chatbot; https://www.heek.com/app/editor)**		

C4. Survey flow



Appendix D.

R code

```
library(MASS)
library(ggpubr)
library(tidyverse)
library(knitr)
library(psych)
library(printr)
library(readr)
library(readxl)
library(psy)
library(dplyr)
library(corpcor)
library(GPArotation)
library(car)
library(mvnormtest)
library(pastecs)
library(reshape)
library(Hmisc)
library(polycor)
library(scales)
library(ggplot2)
library(heplots)

### LOADING DATA
D_chatbots <- read_excel('DATA.xlsx')
View(D_chatbots)

### RESCALING VARIABLES; allows comparability
USQ_total.rescaled <- rescale(D_chatbots$USQ_total)
SUISQ_total.rescaled <- rescale(D_chatbots$SUISQ_total)
GEEK_total.rescaled <- rescale(D_chatbots$GEEK_total)
TD.rescaled <- rescale(D_chatbots$TD)
Familiarity.rescaled <- rescale(D_chatbots$Familiarity)
USQ_24.rescaled <- rescale(D_chatbots$USQ_24_final)

### Check for outliers in relevant variables
outlier_values <- boxplot.stats(D_chatbots$USQ_total)$out   #USQ
boxplot(D_chatbots$USQ_total, boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=1)

outlier_values <- boxplot.stats(D_chatbots$SUISQ_total)$out #SUISQ
boxplot(D_chatbots$SUISQ_total, boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=1)
```

```
outlier_values <- boxplot.stats(D_chatbots$GEEK_total)$out  #GEEK
boxplot(D_chatbots$GEEK_total, boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=1)

outlier_values <- boxplot.stats(D_chatbots$TD)$out   #TD
boxplot(D_chatbots$TD, boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=1)

outlier_values <- boxplot.stats(D_chatbots$Familiarity)$out   #Familiarity
boxplot(D_chatbots$Familiarity, boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=1)

View(outlier_values) ## two outliers. 2 out of 48 participants are extremel
y familiar
```

Part 1: EXPLORATORY ANALYSIS OF USQ

```
### Assumption checks
## 1.correlations between .3 and .9 // at least one >0.3 with another varia
ble #://rstudio-pubs-static.s3.amazonaws.com/363499_73a1c1a94da148b6ad81e6e
b8dc1b771.html & in line with Silderhuis & Borsci (2020)
USQ_only <- D_chatbots[,c(23:64)]
corr_USQ <- cor(USQ_only)
summary(corr_USQ)
summary(USQ_only)

## 2.Kaiser criterion above .5
Data <- data.frame(rbind(USQ_only))
kmo <- function(x){
  x <- subset(x, complete.cases(x))            # Omit missing values
  r <- cor(x)                                  # Correlation matrix
  r2 <- r^2                                 # Squared correlation coefficien
ts
  i <- solve(r)                           # Inverse matrix of correlation matr
ix
  d <- diag(i)                            # Diagonal elements of inverse matri
x
  p2 <- (-i/sqrt(outer(d, d)))^2   # Squared partial correlation coefficien
ts
  diag(r2) <- diag(p2) <- 0                     # Delete diagonal elements
  KMO <- sum(r2)/(sum(r2)+sum(p2))              # Equation for KMO test
  MSA <- colSums(r2)/(colSums(r2)+colSums(p2))  # Equation for individual M
SA

  return(list(KMO=KMO, MSA=MSA))

}

#run the test
kmo(Data)          ### 10 items are below MSA .5

## 3.Barlett's test
bartlett.test(Data)
cortest.bartlett(Data)
```

```
# FA ANALYSIS of all 42 items https://gaopinghuang0.github.io/2018/02/09/exploratory-factor-analysis-notes-and-R-code#what-is-communality
pc1 <- principal(Data, nfactors = 42, rotate = 'none')
pc1

plot(pc1$values, xlab = '# of components', ylab = 'Eigenvalue', type = 'b')
abline(h = 1, col = 'red') # it appears that there are 7 factors (>= Eigenvalue 1)

### Go with all 42 items.
pc2 <- principal(Data, nfactors = 7, rotate = 'none')
pc2

mean(pc2$communality) #h2
mean(pc2$uniquenesses) #u2/noise

## Inspection of the scree plot (elbow criterium) could also indicate 5 factors so check if 5 or 7 is better
pc3 <- principal(Data, nfactors = 5, rotate = 'none')
pc3

mean(pc3$communality) #h2
mean(pc3$uniquenesses) #u2/noise

fa.parallel(Data, fm = 'ml', fa = 'both') # Parallel analysis for additional insight indicates 4-5 factors.

## Parallel analysis suggests that the number of factors =  2  and the number of components =  2

#For now 5 factors seems best considering the proportion explained and parallel analysis shows that 5 is the conservative option
factor.model(pc3$loadings)

residuals <- factor.residuals(corr_USQ, pc3$loadings)
residuals <- as.matrix(residuals[upper.tri(residuals)])

large.resid <- abs(residuals) > 0.05
# proportion of the large residuals
sum(large.resid)/nrow(residuals)
hist(residuals)

#check if <50% of residuals are >.05

#ROTATION; oblimin
pc3_obl <- principal(Data, nfactors = 5, rotate = 'oblimin')
print.psych(pc3_obl, cut = 0.2, scores = T)

pc3_obl$loadings %*% pc3_obl$Phi

FA_structure <- function(fa, cut = 0.2, decimals = 2){
  struc_matrix <- fa.sort(fa$loadings %*% fa$Phi)
  struc_matrix <- data.frame(ifelse(abs(struc_matrix) < cut, '', round(struc_matrix,
decimals)))
 return(struc_matrix)
```

```
}
FA_structure(pc3_obl, cut = 0.3)

## create mutually exclusive factors. >.5 loading only and exclude items if
they load on >2 factors with  >.4 loading

## RELIABILITY ANALYSIS of remaining 23 items
fa1 <- USQ_only[,c(27,24,22,26,29,37,30,28,23,8,9,31)]
fa2 <- USQ_only[,c(2,1,4,5,6)]
fa3 <- USQ_only[,c(20,21,19,13)]
fa4 <- USQ_only[,c(11,10)]
#fa5 <- USQ_only[,c(17)] #only one item, remove this factor?

psych::alpha(fa1)
psych::alpha(fa2)
psych::alpha(fa3)
psych::alpha(fa4)
#psych::alpha(fa5)

### no r.drops <= .3 so no items will be excludedresiduals_4 <- factor.resi
duals(corr_USQ, pc4$loadings)
residuals_4 <- as.matrix(residuals_4[upper.tri(residuals_4)])

large.resid_4 <- abs(residuals_4) > 0.05
# proportion of the large residuals
sum(large.resid_4)/nrow(residuals_4)
hist(residuals_4)

#ROTATION; oblimin
pc4_obl <- principal(Data, nfactors = 4, rotate = 'oblimin')
print.psych(pc4_obl, cut = 0.2, scores = T)

pc4_obl$loadings %*% pc4_obl$Phi

FA_structure <- function(fa, cut = 0.2, decimals = 2){
  struc_matrix <- fa.sort(fa$loadings %*% fa$Phi)
  struc_matrix <- data.frame(ifelse(abs(struc_matrix) < cut, '', round(stru
c_matrix,
decimals)))
 return(struc_matrix)
}
FA_structure(pc4_obl, cut = 0.3)

## create mutually exclusive factors. >.5 loading only and exclude items if
they load on >2 factors with  >.4 loading

#RELIABILITY ANALYSIS of remaining 24 items
fa1b <- USQ_only[,c(24,22,12,26,14,23,10,8,9,31,13,11)]
fa2b <- USQ_only[,c(2,1,4,3,5,6)]
fa3b <- USQ_only[,c(20,21,19,32)]
fa4b <- USQ_only[,c(42,17)]

psych::alpha(fa1b)
psych::alpha(fa2b)
psych::alpha(fa3b)
```

```
psych::alpha(fa4b)
psych::alpha(USQ_24)

## Check is the new data with 24 items meets the criteria
## 1.correlations between .3 and .9 and at least one >.3
USQ_24 <- USQ_only[,c(1,2,3,4,5,6,8,9,10,11,12,13,14,17,19,20,21,22,23,24,2
6,31,32,42)]
corr_USQ_24 <- cor(USQ_24)

summary(corr_USQ_24)
summary(USQ_24)

#run the test
kmo(Data_24)         #all items above MSA .5

## 3.Barlett's test
bartlett.test(Data_24)
cortest.bartlett(Data_24) #significant

### Retained 24 items meet the criteria so 4 factors can be posed
```

PART 2: CORRELATIONAL ANALYSIS BETWEEN USQ AND SUISQ-MR

```
###1. check linearity/correlational relationship between USQ retained 24 it
ems and SUISQ
D_chatbots %>%                    #USQ_24 AND SUISQ
  ggplot(aes(x = USQ_24_final,
             y = SUISQ_total)) +
  geom_point()+
  geom_smooth(method = "lm", se = F, col = 'red')

D_chatbots %>%                    #SUISQ AND TD
  ggplot(aes(x = SUISQ_total,
             y = TD)) +
  geom_point()+
  geom_smooth(method = "lm", se = F, col = 'red')

D_chatbots %>%                    #USQ AND TD
  ggplot(aes(x = USQ_24_final,
             y = TD)) +
  geom_point()+
  geom_smooth(method = "lm", se = F, col = 'red')

###2. check the normality of data
shapiro.test(D_chatbots$USQ_24_final)   #not normally distributed, shapiro
is not significant

shapiro.test(D_chatbots$SUISQ_total)

shapiro.test(D_chatbots$TD)

ggqqplot(D_chatbots$USQ_24_final, main = 'Distribution USQ_24_final', ylab
= 'USQ_24')

ggqqplot(D_chatbots$SUISQ_total, main = 'Distribution SUISQ_total', ylab =
'SUISQ-MR')
```

```
ggqqplot(D_chatbots$TD, main = 'Distribution TD', ylab = 'TD')

qplot(D_chatbots$USQ_24_final,
      geom = 'histogram',
      main = 'Distribution USQ_24',
      xlab = 'USQ_24')

qplot(D_chatbots$SUISQ_total,
      geom = 'histogram',
      main = 'Distribution SUISQ_total',
      xlab = 'SUISQ_total')

qplot(D_chatbots$TD,
      geom = 'histogram',
      main = 'Distribution TD',
      xlab = 'TD')

### data is not normally distributed so Kendall's Tau will be executed
cor.test(D_chatbots$USQ_24_final, D_chatbots$SUISQ_total, method = 'kendall
', exact = F)  # significant

### CI 97.5%
correlation_SUISQ_USQ_24 <- as.data.frame(cbind(USQ_24.rescaled, SUISQ_tota
l.rescaled))

h <- function(d){
  temp <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp$USQ_24.rescaled, temp$SUISQ_total.rescale
d, method = 'kendall', exact = F)$estimate))
}
r_estimate_24 <- replicate(9999, h(correlation_SUISQ_USQ_24))
(CI_est_24 <- quantile(r_estimate_24, c(0.025, 0.975)))

hist(r_estimate_24)

summary(r_estimate_24)

corr_graph_24 <- ggplot(correlation_SUISQ_USQ_24, aes(x = USQ_24.rescaled,
y = SUISQ_total.rescaled), xlab = 'USQ_24_final', ylab = 'SUISQ-MR') + geom
_point() + geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)

corr_graph_24 <- corr_graph_24 + labs(title = 'Correlation USQ_24_final and
SUISQ-MR', x = 'USQ_24_final', y = 'SUISQ-MR')

plot(corr_graph_24)

# Now for every individual factor

### data is not normally distributed so Kendall's Tau will be executed
### FA1B
cor.test(D_chatbots$fa1b, D_chatbots$SUISQ_total, method = 'kendall', exact
= F)  # significant

USQ_fa1b.rescaled <- rescale(D_chatbots$fa1b)
corr_SUISQ_fa1b <- as.data.frame(cbind(USQ_fa1b.rescaled, SUISQ_total.resca
led))
```

```
h1 <- function(d){
  temp1 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp1$USQ_fa1b.rescaled, temp1$SUISQ_total.res
caled, method = 'kendall', exact = F)$estimate))
}
r_estimate_fa1b <- replicate(9999, h1(corr_SUISQ_fa1b))
(CI_est_fa1b <- quantile(r_estimate_fa1b, c(0.025, 0.975)))

hist(r_estimate_fa1b)
summary(r_estimate_fa1b)

corr_graph_fa1b <- ggplot(corr_SUISQ_fa1b, aes(x = USQ_fa1b.rescaled, y = S
UISQ_total.rescaled), xlab = 'USQ_fa1b', ylab = 'SUISQ-MR') + geom_point()
+
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_fa1b <- corr_graph_fa1b + labs(title = 'Correlation USQ_fa1b and
SUISQ-MR', x = 'USQ_fa1b', y = 'SUISQ-MR')

plot(corr_graph_fa1b)

### fa2b
cor.test(D_chatbots$fa2b, D_chatbots$SUISQ_total, method = 'kendall', exact
= F)  # significant

USQ_fa2b.rescaled <- rescale(D_chatbots$fa2b)
corr_SUISQ_fa2b <- as.data.frame(cbind(USQ_fa2b.rescaled, SUISQ_total.resca
led))

h2 <- function(d){
  temp2 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp2$USQ_fa2b.rescaled, temp2$SUISQ_total.res
caled, method = 'kendall', exact = F)$estimate))
}
r_estimate_fa2b <- replicate(9999, h2(corr_SUISQ_fa2b))
(CI_est_fa2b <- quantile(r_estimate_fa2b, c(0.025, 0.975)))

hist(r_estimate_fa2b)
summary(r_estimate_fa2b)

corr_graph_fa2b <- ggplot(corr_SUISQ_fa2b, aes(x = USQ_fa2b.rescaled, y = S
UISQ_total.rescaled), xlab = 'USQ_fa2b', ylab = 'SUISQ-MR') + geom_point()
+
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_fa2b <- corr_graph_fa2b + labs(title = 'Correlation USQ_fa2b and
SUISQ-MR', x = 'USQ_fa2b', y = 'SUISQ-MR')

plot(corr_graph_fa2b)

### fa3b
cor.test(D_chatbots$fa3b, D_chatbots$SUISQ_total, method = 'kendall', exact
= F)  # NOT significant

### fa4b
cor.test(D_chatbots$fa4b, D_chatbots$SUISQ_total, method = 'kendall', exact
= F)  # NOT significant
```

```
### Comparisons with factors of SUISQ-MR
SUISQ_fa1.rescaled <- rescale(D_chatbots$fa1_SUISQ)
SUISQ_fa2.rescaled <- rescale(D_chatbots$fa2_SUISQ)
SUISQ_fa3.rescaled <- rescale(D_chatbots$fa3_SUISQ)
SUISQ_fa4.rescaled <- rescale(D_chatbots$fa4_SUISQ)
###FA1B USQ
#SUISQ_1
cor.test(D_chatbots$fa1b, D_chatbots$fa1_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ1_fa1b <- as.data.frame(cbind(USQ_fa1b.rescaled, SUISQ_fa1.rescal
ed))

s1 <- function(d){
  temp3 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp3$USQ_fa1b.rescaled, temp3$SUISQ_fa1.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ1_fa1b <- replicate(9999, s1(corr_SUISQ1_fa1b))
(CI_est_SUISQ1_fa1b <- quantile(r_estimate_SUISQ1_fa1b, c(0.025, 0.975)))

hist(r_estimate_SUISQ1_fa1b)
summary(r_estimate_SUISQ1_fa1b)

corr_graph_SUISQ1_fa1b <- ggplot(corr_SUISQ1_fa1b, aes(x = USQ_fa1b.rescale
d, y = SUISQ_fa1.rescaled), xlab = 'USQ_fa1b', ylab = 'SUISQ-MR_fa1') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ1_fa1b <- corr_graph_SUISQ1_fa1b + labs(title = 'Correlatio
n USQ_fa1b and SUISQ-MR_fa1', x = 'USQ_fa1b', y = 'SUISQ-MR_fa1')

plot(corr_graph_SUISQ1_fa1b)

#SUISQ_2
cor.test(D_chatbots$fa1b, D_chatbots$fa2_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ2_fa1b <- as.data.frame(cbind(USQ_fa1b.rescaled, SUISQ_fa2.rescal
ed))

s2 <- function(d){
  temp4 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp4$USQ_fa1b.rescaled, temp4$SUISQ_fa2.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ2_fa1b <- replicate(9999, s2(corr_SUISQ2_fa1b))
(CI_est_SUISQ2_fa1b <- quantile(r_estimate_SUISQ2_fa1b, c(0.025, 0.975)))

hist(r_estimate_SUISQ2_fa1b)
summary(r_estimate_SUISQ2_fa1b)

corr_graph_SUISQ2_fa1b <- ggplot(corr_SUISQ2_fa1b, aes(x = USQ_fa1b.rescale
d, y = SUISQ_fa2.rescaled), xlab = 'USQ_fa1b', ylab = 'SUISQ-MR_fa2') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
```

```
corr_graph_SUISQ2_fa1b <- corr_graph_SUISQ2_fa1b + labs(title = 'Correlatio
n USQ_fa1b and SUISQ-MR_fa2', x = 'USQ_fa1b', y = 'SUISQ-MR_fa2')

plot(corr_graph_SUISQ2_fa1b)

#SUISQ_3
cor.test(D_chatbots$fa1b, D_chatbots$fa3_SUISQ, method = 'kendall', exact =
F)  # NOT significant

#SUISQ_4
cor.test(D_chatbots$fa1b, D_chatbots$fa4_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ4_fa1b <- as.data.frame(cbind(USQ_fa1b.rescaled, SUISQ_fa4.rescal
ed))

s3 <- function(d){
  temp5 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp5$USQ_fa1b.rescaled, temp5$SUISQ_fa4.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ4_fa1b <- replicate(9999, s3(corr_SUISQ4_fa1b))
(CI_est_SUISQ4_fa1b <- quantile(r_estimate_SUISQ4_fa1b, c(0.025, 0.975)))

hist(r_estimate_SUISQ4_fa1b)summary(r_estimate_SUISQ4_fa1b)
summary(r_estimate_SUISQ4_fa1b)

corr_graph_SUISQ4_fa1b <- ggplot(corr_SUISQ4_fa1b, aes(x = USQ_fa1b.rescale
d, y = SUISQ_fa4.rescaled), xlab = 'USQ_fa1b', ylab = 'SUISQ-MR_fa4') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ4_fa1b <- corr_graph_SUISQ4_fa1b + labs(title = 'Correlatio
n USQ_fa1b and SUISQ-MR_fa4', x = 'USQ_fa1b', y = 'SUISQ-MR_fa4')

plot(corr_graph_SUISQ4_fa1b)

###FA2B USQ
#SUISQ_1
cor.test(D_chatbots$fa2b, D_chatbots$fa1_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ1_fa2b <- as.data.frame(cbind(USQ_fa2b.rescaled, SUISQ_fa1.rescal
ed))

s4 <- function(d){
  temp6 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp6$USQ_fa2b.rescaled, temp6$SUISQ_fa1.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ1_fa2b <- replicate(9999, s4(corr_SUISQ1_fa2b))
(CI_est_SUISQ1_fa2b <- quantile(r_estimate_SUISQ1_fa2b, c(0.025, 0.975)))

hist(r_estimate_SUISQ1_fa2b)
summary(r_estimate_SUISQ1_fa2b)
```

```
corr_graph_SUISQ1_fa2b <- ggplot(corr_SUISQ1_fa2b, aes(x = USQ_fa2b.rescale
d, y = SUISQ_fa1.rescaled), xlab = 'USQ_fa2b', ylab = 'SUISQ-MR_fa1') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ1_fa2b <- corr_graph_SUISQ1_fa2b + labs(title = 'Correlatio
n USQ_fa2b and SUISQ-MR_fa1', x = 'USQ_fa2b', y = 'SUISQ-MR_fa1')

plot(corr_graph_SUISQ1_fa2b)

#SUISQ_2
cor.test(D_chatbots$fa2b, D_chatbots$fa2_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ2_fa2b <- as.data.frame(cbind(USQ_fa2b.rescaled, SUISQ_fa2.rescal
ed))

s5 <- function(d){
  temp7 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp7$USQ_fa2b.rescaled, temp7$SUISQ_fa2.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ2_fa2b <- replicate(9999, s5(corr_SUISQ2_fa2b))
(CI_est_SUISQ2_fa2b <- quantile(r_estimate_SUISQ2_fa2b, c(0.025, 0.975)))

hist(r_estimate_SUISQ2_fa2b)
summary(r_estimate_SUISQ2_fa2b)

corr_graph_SUISQ2_fa2b <- ggplot(corr_SUISQ2_fa2b, aes(x = USQ_fa2b.rescale
d, y = SUISQ_fa2.rescaled), xlab = 'USQ_fa2b', ylab = 'SUISQ-MR_fa2') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ2_fa2b <- corr_graph_SUISQ2_fa2b + labs(title = 'Correlatio
n USQ_fa2b and SUISQ-MR_fa2', x = 'USQ_fa2b', y = 'SUISQ-MR_fa2')

plot(corr_graph_SUISQ2_fa2b)

#SUISQ_3
cor.test(D_chatbots$fa2b, D_chatbots$fa3_SUISQ, method = 'kendall', exact =
F)  # NOT significant

#SUISQ_4
cor.test(D_chatbots$fa2b, D_chatbots$fa4_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ4_fa2b <- as.data.frame(cbind(USQ_fa2b.rescaled, SUISQ_fa4.rescal
ed))

s6 <- function(d){
  temp8 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp8$USQ_fa2b.rescaled, temp8$SUISQ_fa4.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ4_fa2b <- replicate(9999, s6(corr_SUISQ4_fa2b))
(CI_est_SUISQ4_fa2b <- quantile(r_estimate_SUISQ4_fa2b, c(0.025, 0.975)))
```

```
hist(r_estimate_SUISQ4_fa2b)
summary(r_estimate_SUISQ4_fa2b)

corr_graph_SUISQ4_fa2b <- ggplot(corr_SUISQ4_fa2b, aes(x = USQ_fa2b.rescale
d, y = SUISQ_fa4.rescaled), xlab = 'USQ_fa2b', ylab = 'SUISQ-MR_fa4') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ4_fa2b <- corr_graph_SUISQ4_fa2b + labs(title = 'Correlatio
n USQ_fa2b and SUISQ-MR_fa4', x = 'USQ_fa2b', y = 'SUISQ-MR_fa4')

plot(corr_graph_SUISQ4_fa2b)

###FA3B USQ
USQ_fa3b.rescaled <- rescale(D_chatbots$fa3b)
#SUISQ_1
cor.test(D_chatbots$fa3b, D_chatbots$fa1_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ1_fa3b <- as.data.frame(cbind(USQ_fa3b.rescaled, SUISQ_fa1.rescal
ed))

s7 <- function(d){
  temp9 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp9$USQ_fa3b.rescaled, temp9$SUISQ_fa1.resca
led, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ1_fa3b <- replicate(9999, s7(corr_SUISQ1_fa3b))
(CI_est_SUISQ1_fa3b <- quantile(r_estimate_SUISQ1_fa3b, c(0.025, 0.975)))

hist(r_estimate_SUISQ1_fa3b)
summary(r_estimate_SUISQ1_fa3b)

corr_graph_SUISQ1_fa3b <- ggplot(corr_SUISQ1_fa3b, aes(x = USQ_fa3b.rescale
d, y = SUISQ_fa1.rescaled), xlab = 'USQ_fa3b', ylab = 'SUISQ-MR_fa1') + geo
m_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ1_fa3b <- corr_graph_SUISQ1_fa3b + labs(title = 'Correlatio
n USQ_fa3b and SUISQ-MR_fa1', x = 'USQ_fa3b', y = 'SUISQ-MR_fa1')

plot(corr_graph_SUISQ1_fa3b)

#SUISQ_2
cor.test(D_chatbots$fa3b, D_chatbots$fa2_SUISQ, method = 'kendall', exact =
F)  # significant

corr_SUISQ2_fa3b <- as.data.frame(cbind(USQ_fa3b.rescaled, SUISQ_fa2.rescal
ed))

s8 <- function(d){
  temp10 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp10$USQ_fa3b.rescaled, temp10$SUISQ_fa2.res
caled, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ2_fa3b <- replicate(9999, s8(corr_SUISQ2_fa3b))
(CI_est_SUISQ2_fa3b <- quantile(r_estimate_SUISQ2_fa3b, c(0.025, 0.975)))
```

```r
hist(r_estimate_SUISQ2_fa3b)
summary(r_estimate_SUISQ2_fa3b)

corr_graph_SUISQ2_fa3b <- ggplot(corr_SUISQ2_fa3b, aes(x = USQ_fa3b.rescaled, y = SUISQ_fa2.rescaled), xlab = 'USQ_fa3b', ylab = 'SUISQ-MR_fa2') + geom_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ2_fa3b <- corr_graph_SUISQ2_fa3b + labs(title = 'Correlation USQ_fa3b and SUISQ-MR_fa2', x = 'USQ_fa3b', y = 'SUISQ-MR_fa2')

plot(corr_graph_SUISQ2_fa3b)

#SUISQ_3
cor.test(D_chatbots$fa3b, D_chatbots$fa3_SUISQ, method = 'kendall', exact = F)  # NOT significant

#SUISQ_4
cor.test(D_chatbots$fa3b, D_chatbots$fa4_SUISQ, method = 'kendall', exact = F)  # NOT significant

###FA4B USQ
USQ_fa4b.rescaled <- rescale(D_chatbots$fa4b)
#SUISQ_1
cor.test(D_chatbots$fa4b, D_chatbots$fa1_SUISQ, method = 'kendall', exact = F)  # NOT significant

#SUISQ_2
cor.test(D_chatbots$fa4b, D_chatbots$fa2_SUISQ, method = 'kendall', exact = F)  # significant

corr_SUISQ2_fa4b <- as.data.frame(cbind(USQ_fa4b.rescaled, SUISQ_fa2.rescaled))

s10 <- function(d){
  temp12 <- d[sample(nrow(d), replace = T), ]
  return(as.numeric(cor.test(temp12$USQ_fa4b.rescaled, temp12$SUISQ_fa2.rescaled, method = 'kendall', exact = F)$estimate))
}
r_estimate_SUISQ2_fa4b <- replicate(9999, s10(corr_SUISQ2_fa4b))
(CI_est_SUISQ2_fa4b <- quantile(r_estimate_SUISQ2_fa4b, c(0.025, 0.975)))

hist(r_estimate_SUISQ2_fa4b)
summary(r_estimate_SUISQ2_fa4b)

corr_graph_SUISQ2_fa4b <- ggplot(corr_SUISQ2_fa4b, aes(x = USQ_fa4b.rescaled, y = SUISQ_fa2.rescaled), xlab = 'USQ_fa4b', ylab = 'SUISQ-MR_fa2') + geom_point() +
geom_smooth(method = 'lm', color = 'red', se = T, level = 0.975)
corr_graph_SUISQ2_fa4b <- corr_graph_SUISQ2_fa4b + labs(title = 'Correlation USQ_fa4b and SUISQ-MR_fa2', x = 'USQ_fa4b', y = 'SUISQ-MR_fa2')

plot(corr_graph_SUISQ2_fa4b)

#SUISQ_3
cor.test(D_chatbots$fa4b, D_chatbots$fa3_SUISQ, method = 'kendall', exact = F)  # NOT significant
```

```
#SUISQ_4
cor.test(D_chatbots$fa4b, D_chatbots$fa4_SUISQ, method = 'kendall', exact =
F)  # NOT significant

### Comparisons between total 24-USQ and individual factors of SUISQ_MR
#F1
cor.test(D_chatbots$USQ_24_final, D_chatbots$fa1_SUISQ, method = 'kendall',
exact = F)  # significant
#F2
cor.test(D_chatbots$USQ_24_final, D_chatbots$fa2_SUISQ, method = 'kendall',
exact = F)  # significant

#F3
cor.test(D_chatbots$USQ_24_final, D_chatbots$fa3_SUISQ, method = 'kendall',
exact = F)  # NOT significant

#F4
cor.test(D_chatbots$USQ_24_final, D_chatbots$fa4_SUISQ, method = 'kendall',
exact = F)  # significant
```

PART 3: NON-LINEAR REGRESSIONS FAMILIARITY AND GEEKISM - USQ scores

```
#LINEAR REGRESSION BETWEEN FAMILIARITY AND TOTAL USQ SCORES
# http://www.sthda.com/english/articles/40-regression-analysis/167-simple-l
inear-regression-in-r/
ggplot(D_chatbots, aes(x = AVG_Fam, y = AVG_24_score)) +
  geom_point() +
  geom_smooth(se = F, method = 'lm', col = 'red')

cor(D_chatbots$AVG_Fam, D_chatbots$AVG_24_score)

m_fam <- lm(AVG_24_score ~ AVG_Fam, data = D_chatbots)
summary(m_fam)
anova(m_fam)

library(rstanarm)
options(mc.cores = 2)

M_1 <- stan_glm(AVG_24_score ~ 1 + AVG_Fam,
            data = D_chatbots)

save(M_1, D_chatbots, file = 'M_1.Rda')

load('M_1.Rda')
fixef(M_1)

library(bayr)
coef(M_1)

USQ_pred_M_1 <-
   D_chatbots %>%
   mutate(pred_M_1 = predict(M_1)$center,
                     err_M_1 = AVG_24_score - pred_M_1)
```

```
USQ_pred_M_1 %>%
   ggplot(aes(x = AVG_Fam, y = pred_M_1)) +
   geom_point()

USQ_pred_M_1 %>%
   ggplot(aes(x = err_M_1)) +
   geom_histogram()

library(quantreg)
USQ_pred_M_1 %>%
   ggplot(aes(x = pred_M_1,
              y = err_M_1)) +
   geom_point() +
   geom_quantile()

# assumptions not met.


#LINEAR REGRESSION BETWEEN GEEKISM AND TOTAL USQ SCORES

ggplot(D_chatbots, aes(x = AVG_Geekism, y = AVG_24_score)) +
  geom_point() +
  geom_smooth(se = F, method = 'lm', col = 'red')

cor(D_chatbots$AVG_Geekism, D_chatbots$AVG_24_score)

m_geek <- lm(AVG_24_score ~ AVG_Geekism, data = D_chatbots)
summary(m_geek)
anova(m_geek)

library(rstanarm)
options(mc.cores = 2)

M_2 <- stan_glm(AVG_24_score ~ 1 + AVG_Geekism,
            data = D_chatbots)

save(M_2, D_chatbots, file = 'M_2.Rda')

load('M_2.Rda')
fixef(M_2)

library(bayr)
coef(M_2)

USQ_pred_M_2 <-
   D_chatbots %>%
   mutate(pred_M_2 = predict (M_2)$center,
                     err_M_2 = AVG_24_score - pred_M_2)

USQ_pred_M_2 %>%
   ggplot(aes(x = AVG_Geekism, y = pred_M_2)) +
   geom_point()

USQ_pred_M_2 %>%
   ggplot(aes(x = err_M_2)) +
   geom_histogram()
```

```
library(quantreg)
USQ_pred_M_2 %>%
   ggplot(aes(x = pred_M_2,
              y = err_M_2)) +
   geom_point() +
   geom_quantile()

# assumptions not met.

#NON-LINEAR REGRESSION

###FAM loess nonlinear least squares regression
ggplot(D_chatbots, aes(x = AVG_Fam, y = AVG_24_score)) +
  geom_point() +
  geom_smooth(se = T, method = 'loess', col = 'red')

m_fam_2 <- loess(AVG_24_score ~ AVG_Fam, data = D_chatbots)
summary(m_fam_2)

library(mgcv)
gam_mod1 <- gam(AVG_24_score ~ s(AVG_Fam, k = 5), data = D_chatbots)
plot(gam_mod1)
coef(gam_mod1)
anova(gam_mod1) ### Significant. Nonlinear is best fit

library(devtools)
devtools::install_github('ProcessMiner/nlcor')
library(nlcor)
c_fam <- nclor(D_chatbots$AVG_Fam, D_chatbots$AVG_24_score, refine = 0.5,
      plt = T)
c_fam$cor.estimate
c_fam$adjusted.p.value
print(c_fam$cor.plot)



###GEEKISM loess nonlinear least squares regression
ggplot(D_chatbots, aes(x = AVG_Geekism, y = AVG_24_score)) +
  geom_point() +
  geom_smooth(se = T, method = 'loess', col = 'red')

m_geek_2 <- loess(AVG_24_score ~ AVG_Geekism, data = D_chatbots)
summary(m_geek_2)

library(mgcv)
gam_mod2 <- gam(AVG_24_score ~ s(AVG_Geekism, k = 5), data = D_chatbots)
plot(gam_mod2)
coef(gam_mod2)
anova(gam_mod2) ### Not Significant. Neither linear/nonlinear fit

library(nlcor)
c_geek <- nclor(D_chatbots$AVG_Geekism, D_chatbots$AVG_24_score,
      refine = 0.5, plt = T)
c_geek$cor.estimate
c_geek$adjusted.p.value
print(c_geek$cor.plot)
```