Development of a Preliminary Measurement Tool of User Satisfaction for Information-Retrieval Chatbots

Bachelor Thesis

Lisa Waldera

Dept. of Cognitive Psychology and Ergonomics

University of Twente

Examination Committee:

Dr. Simone Borsci

Dr. Martin Schmettow

16.08.2019

Abstract

Based on the findings of Tariverdiyeva and Borsci (2019), it can be concluded that the implementation of 18 key features of information-retrieval chatbots are directly leading to a change in perceived quality and user satisfaction. Standard measurement tools for the usability of software, like UMUX-Lite (Lewis et al., 2013), are found to be failing at capturing these specific requirements for chatbots. This research aims to reassess and extend the list of relevant key features found by Tariverdiyeva and Borsci (2019). Moreover, this research aims to develop a preliminary list of questionnaire items that can be used in a future measurement tool of user satisfaction for information-retrieval chatbots from an end user's perspective. This is achieved by conducting a systematic literature review and discussions with possible end users in focus groups in the first experimental phase and conducting a usability testing with different chatbots in the second experimental phase. After the first phase, a total of 14 features were found to be relevant to measure user satisfaction. Questionnaire items for these features were generated and used for the development of a preliminary questionnaire – UMT/C. After the second experimental phase, the UMT/C was found to be correlating with the standard measurement tool UMUX-Lite. A factor analysis was applied to the questionnaire data which showed a total of 9 factors being measured by 25 items. While some of the factors correlated highly with the results of the UMUX-Lite scale, others were found to capture additional information related to user satisfaction. Therefore, the present study recommends to test the quality of this preliminary questionnaire and to refine the pool of items in order to develop a valid and reliable tool for assessing the quality of chatbots from an end user's perspective.

Keywords: information-retrieval chatbots, usability testing, UMUX-Lite, questionnaire development, factor analysis

Table of contents

Introduction

Enabled by the rise of artificial intelligence, increased interconnectivity and improved language procession, the implementation of dialogue-based systems has been growing increasingly during the last few years (McTear, 2017). As a consequence, people are more likely to be confronted with conversational software when operating services and requesting help online.

The beginnings of this development, however, can be traced back to the first application that allowed a conversation with a machine via written language – Eliza. The conversational software Eliza was developed in the early 1960s and used template-based responses to imitate a non-directional conversation style (Dale, 2016). It was found that the software was able to trick the user into thinking it was a person he or she was conversing with. This success sparked the interest of researchers to develop conversational software that could even pass the Turing test (Dale, 2016).

Nowadays, dialogue-based systems can be found in a variety of implementations and types. Virtual assistants or conversational agents can be divided in four main categories: spoken dialogue systems (SDSs), voice user interface (VUIs), embodied conversational agents (ECA) and chat robots (short: chatbot). All of them differ in their specific goals and operational methods (McTear, Callejas, & Barres, 2016). Chatbots are allowing the user to engage in a conversational dialogue with the software. Natural language in form of speech or text, is used to communicate the user's goal to the chatbot (Dale, 2016).

The concept of text-based conversation can be found in numerous aspects of our lives. According to Dale (2016), 49% of the 18-29-year-olds and 37% of the 30-49-year-olds are using messaging apps. For example, Facebook messenger alone has one billion users worldwide. Naturally, communicating via short messages and typed interactions are becoming almost essential in our daily conversations (Dale, 2016). Thus, the text-based conversation with a chatbot should feel familiar and effortless for the experienced user.

But simple familiarity with the conversational style is not the reason why people are particularly motivated to engage with chatbots. A study from Brandtzaeg and Følstad (2017) showed that the most named benefit of using this conversational technology is increased productivity. Quick responses and immediate feedback draw people towards chatbot services. Additionally, people expect some level of entertainment and sociality of the program as long as it

is not hindering the efficiency to quickly reach the intended goal (Brandtzaeg & Følstad, 2017).

Due to open source code and available development platforms for chatbots online, it is simpler and faster than ever to create and launch conversational agents (Radziwill & Benton, 2017). While chatbots are implemented increasingly, there is little evidence of the impact of design features on the assessment of quality by possible end users. In order to assure a certain level of quality of chatbots, Radziwill and Benton (2017) propose a quality assessment procedure for chatbot implementation. They suggest the Analytic Hierarchy Process (AHP) approach but this procedure would allow researchers to compare the quality of only two chatbots at the same time (Radziwill & Benton, 2017). Manual testing on the other hand is found to be extremely time-consuming and often inaccurate when measuring the quality of interaction with a chatbot (Vasconcelos, Candello, Pinhanez, & dos Santos, 2017).

Thus, a standardized procedure for assessing the quality of interaction with chatbots would enable researchers and designers to quickly draw conclusions about its quality and possible success when operating online. More recent research by Tariverdiyeva and Borsci (2019) laid the foundation for developing a usability measurement tool for information-retrieval chatbots. A three-phase experimental model was implemented. Through literature review, conducting an online survey and interaction tests with chatbots, a first preliminary framework was created to test the usability of chatbots from the end user's perspective. The research indicates that the implementation and quality of certain features (previously referred to as factors) leads to direct changes in user satisfaction. In total, 18 features were identified to be relevant when measuring usability from an end user's perspective.

Moreover, the research study of Tariverdiyeva and Borsci (2019) found that standard tools for measuring usability of systems, like UMUX-Lite (Lewis et al., 2013), exist but fail to capture all of the relevant features when applied to chatbots. The findings of this study suggest that a future measurement tool for chatbots should include the 18 relevant features in order to assess usability of chatbots from an end user's perspective (International Organization for Standardization [9241-11], 2018).

Study aims and Outline

The present study will try to extend the preliminary findings of Tariverdiyeva and Borsci (2019) on the relevance of key features. It aims to confirm the relevance of the 18 features that

were identified to influence user satisfaction for information-retrieval chatbots. Also, it aims to create a preliminary version of a new usability measurement tool for chatbots. This is achieved through the implementation of a two-phase experimental structure.

A systematic literature review and focus group discussions with possible end users are conducted in the first phase of this research. It is expected that the collection relevant features can be confirmed and possibly extended by this research.

RQ1: What features and items are found to be relevant from an end user's perspective when measuring the level of user satisfaction of information-retrieval chatbots?

Moreover, this study will further the previous findings in the second experimental phase. By creating and testing a preliminary list of questionnaire items, these could be used in a future quality measurement tool for chatbots. The second phase aims to test and explore the properties of a preliminary questionnaire to assess people's satisfaction during the simulated interaction with real chatbot services. Data gathered during this usability testing will be used to explore the main underlying factors of the new tool – named Usability Measurement Tool for Chatbots (UMT/C) – and its correlation with a standardized scale of satisfaction – UMUX-Lite (Lewis et al., 2013). The findings of the second experimental phase are expected to enable the creation of a preliminary version of the UMT/C.

RQ2: What questionnaire items should be included in the new measurement tool from an end user's perspective?

Pre-experimental Phase

During the pre-experimental phase, the research group of the University of Twente reviewed the previously done research in the field of chatbots. The findings of the study conducted by Tariverdiyeva and Borsci (2019) were used as a basis and starting point for discussions about the number and relevance of features influencing user satisfaction for chatbots. The initial list contained 18 features (see Table 1) that possible end users marked as directly correlating with the perceived user satisfaction.

Feature name	Description
Response Time	Ability of the chatbot to respond timely to users' requests
Graceful responses in unexpected situations	Ability of the chatbot to gracefully handle unexpected input, communication mismatch and broken line of conversation
Maxim of quantity	Ability of the chatbot to respond in an informative way without adding too much information
Recognition and facilitation of user's goal and intent	Ability of the chatbot to recognize user's intent and guide the user to its goal
Maxim of quality	Ability of the chatbot to avoid false statements/information
Perceived Ease of Use	The degree to which a person believes that to interact with a chatbot would be free of effort
Maxim of manners	Ability of the chatbot to make it is purpose clear and communicate without ambiguity
Engage in on-the-fly problem solving	Ability of the chatbot to solve problems instantly on the spot
Maxim of relation	Ability of the chatbot to provide the relevant and appropriate contribution to people needs at each stage
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
Appropriate degrees of formality	Ability of the chatbot to use appropriate language style for the context
User's privacy and ethical decision making	Ability of the chatbot to protect user's privacy and make ethically appropriate decisions on behalf of the user
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

Meets neurodiversity needs	Ability of the chatbot to meet needs of users independently form their health conditions, well-being, age etc.
Integration with the website	Position in the website and visibility of the chatbot (all pages/specific pages, floating window/pull-out tab/embedded etc.)
Trustworthiness	Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage
Process facilitation and follow up	Ability of the chatbot to inform and update users about the status of their task in progress
Flexibility of linguistic input	*No description*

Table 1: Initial list of features according to Tariverdiyeva and Borsci (2019)

The feature *Flexibility of linguistic input* was retrieved from conversations with possible end users later on in the previous research study. Thus, a group of experts for this research discussed the found aspects of this feature and created a description that was used for further testing: '*How easily the chatbot understands the user's input, regardless of the phrasing*'.

For this research, a systematic literature review was conducted by the research team with different search terms related to quality and usability measurement and features of chatbots. As a result, studies suggested the possible influence and thus, the addition of features to the initial list. The group of researchers discussed these features on the basis of the evidence presented in the research studies and possible relevance for measuring the usability of chatbots from an end user's perspective. Consequently, three features were decided to be added and considered for the following experimental phases:

- Expectation Setting (Luger and Sellen, 2016)
  Description: Make purpose clear, show user what it can and cannot do with chatbot, was taken from maxim of manners
- Personality (Chaves and Gerosa, 2019; Jain, Kumar, Kota and Patel, 2018)
  Description: The chatbot appears to have a (human-like) personality
- Enjoyment (Jain et al., 2018)

Description: How enjoyable the interaction with the chatbot appears to be to the user

Moreover, the research group of experts discussed the descriptions and names of the features found by Tariverdiyeva and Borsci (2019). The following names of features were found to be unclear by the research group and thus decided to be renamed to avoid misunderstandings for participants of this research:

- Perceived credibility *(before: Maxim of quality)*
- Understandability *(before: Maxim of manners)*
- Reference to service *(before: Reference to what is on the screen)*
- Visibility *(before: Integration with the website)*
- User's privacy *(before: User's privacy and ethical decision making)*

The feature *Meets diversity needs* was removed from the pool because a single user cannot be able to measure this feature for every other possible user. It has to be noted that this feature might increase the usability of a chatbot for certain users and thus, should be added to a possible list of relevant features for chatbot designers.

Factor Number	**Factor Name**
F1	Response Time
F2	Engage in on-the-fly problem solving
F3	Trust (general)
F4	Privacy & Security
F5	Perceived credibility
F6	Understandability
F7	Maxim of relation
F8	Appropriate language style
F9	Ability to maintain themed discussion
F10	Maxim of quantity
F11	Ease of use (general)
F12	Flexibility of linguistic input

F13	Visibility (website only)
F14	Ease of starting a conversation
F15	Expectation setting
F16	Reference to service
F17	Process tracking
F18	Recognition and facilitation of user's goal & intent
F19	Graceful responses in unexpected situations
F20	Personality
F21	Enjoyment

Table 2: Preliminary list of features after expert review during the pre-experimental phase

As a result, the final list of features (see Table 2) included 21 features possibly correlating with user satisfaction of chatbots. *Trust* and *Ease of use* were expected to be general features that cannot be seen separately but are included in the other aspects of usability.

In order to develop a measuring tool to test the usability of chatbots, 6 questionnaire items were generated per feature. Each supposedly measuring one individual feature only. The quality of the items was being evaluated by assessing the quality by the research group. As a result, the number of items per feature was reduced to three. This led to an initial item pool of 63 possible items to be tested further in this research.

First Experimental Phase (Focus Group)

Methods

Participants.

The first experimental phase included the set-up of a focus group with possible end users. 16 (n=16) students of the University of Twente participated in the discussions. From which eight identified as female and eight as male. The age of the participants ranged from 19 to 30 with an approximate average age of 22 years. (M=22.1 years, SD=2.84 years). The students were able to sign-up voluntarily on SONA, a recruitment page of the University of Twente, and furthermore

collected through convenience sampling. The nationalities of the participants were stated as German (n=6), Indian (n=5), Bulgarian (n=3) and Dutch (n=2).

The population for this experiment consisted of a variety of educational backgrounds: Psychology (n=10), Communication Science (n=1) and other technical studies (n=5) students.

Material.

An Informed Consent Form and a Participant Information Sheet (PIS) (see Appendix A) were created for the focus groups. A demographic questionnaire (see Appendix A) was created to capture the participants age, gender, nationality and field of study. Moreover, the participants were asked to fill in if they had used a chatbot before (Yes/No). In a next step, the questionnaire asked how often they use it (on a scale of 1 to 5) and how their previous experience with chatbot would be rated (on a scale 1 to 5).

The focus group session was guided by one of three moderators of the research team who followed closely the instructions given on the script (see Appendix A). By generating and using main questions and possible prompts, it could be guaranteed that the discussion panels worked towards similar research goals to ensure possible comparisons of the results. Moreover, all sessions were video recorded using a GoPro Hero 5 model on a table tripod.

In order to gather the data of the end users, a list of the key features and their descriptions (see Appendix A) and of the generated questionnaire items (see Appendix A) were generated.

Procedure.

Participants were given the Participant Information Sheet (PIS) with the Informed consent form. The order of presented items on the working sheets was randomized before every session. The PowerPoint presentation was started.

At three out of four sessions, the moderator was accompanied by an assistant from the research team. The assistant was responsible for distributing and collecting the papers and the proper video recording.

The participants received overview of procedure of the session, followed by a brief introduction of the team. After giving consent, the moderator specifically highlighted the aspect

of the video recording again. If participants did not agree with the recording, the research team would have offered to audio record or take notes instead. The moderator explained the aims of the overall research and the nature of chatbots. This was followed by the distribution of the Demographics questionnaire. Participants were asked to fill in their personal data and hand it back to the present assistant.

The participants received information about general discussion guidelines that would ensure a high quality of information exchange and limit the influence of groupthink.

A short demonstration of the Facebook Messenger chatbot used by the Finnish Airline "Finnair" was presented to familiarize the students with chatbot interactions.

The first task requested the possible end users to evaluate the quality and understandability of features and their descriptions on paper. Moreover, the relevance for testing user satisfaction was noted on the sheet. The research members then started a discussion about the main findings and opinions of the participants.

The second task required the participants to read through the list of items for a questionnaire and to evaluate the quality of them. Additionally, the items were matched to the feature that they were supposedly measuring according to the group. Data was captured by every student on a paper sheet. A short discussion afterwards highlighted the most common opinions and disagreements between the participants.

In the end, the moderator concluded the session. She was offering to answer any questions or provided the participants with the possibility to contact the research team.

Data Analysis.

The video recordings were reviewed by the group of researchers. Information was gathered about the features named in discussions and the end user's opinion on them. Main findings were collected and taken into consideration for the further development of a measurement tool.

Additionally, comments and notes left on the lists of features and items were used as complementing data in order to evaluate the relevance of features.

All participants of the focus groups were asked to rate the relevance of each feature for the usability of a chatbot. The gathered data was analyzed with two scoring systems developed

by members of the research team. Both were used with equal weight in order to evaluate the given answers.

The first scoring system (Score 1) distinguished two conditions: an apparent positive response was coded as “1” and every other as negative, thus “0”. Afterwards, the results were converted into percentages.

The second scoring system (Score 2) refined the conditions and distinguished 4 possible types of answers: very relevant (+1.5), relevant (+1), medium/unsure (-0.5) and irrelevant (-1). The assigned scores in points of every participant for every individual feature were added up.

Both scoring systems were used equally to compare the perceived relevance of the individual features. The research group set 75% (or 12 points) as the direct inclusion criteria for features. Furthermore, the quantitative data was given weight when the scoring systems did not give clear and corresponding results or the score was below the inclusion criteria.

Results

Demographic Analysis.

The 78.5% of the students (n=11) had used a chatbot before, whereas 21.5% (n=3) had not. The participants who previously answered to have used a chatbot provided information about how often they used them and the rating of their experiences (both on a Likert scale). On average, the times of use was rated between “Rarely” and “Sometimes” (M=2.5). The rating of previous experience with chatbots had an average of 3.25, which translated into “Fair”.

Main findings for features.

Both previously described scoring systems were applied to the data of possibly relevant features for information-retrieval chatbots (see Table 3).

Feature No.	Score 1 (in %)	Score 2 (in points)
F5	100	17

F6	100	16.5
F10	100	16.5
F11	100	14.5
F15	100	17
F1	93.75	14.5
F12	93.75	15
F16	93.75	9.5
F4	87.5	13
F9	87.5	14.5
F13	87.5	12.5
F18	87.5	14
F3	81.25	9
F7	81.75	12.5
F8	75	10
F17	75	9
F2	68.75	8
F14	68.75	6.5
F19	68.75	7.5
F20	50	0.5
F21	50	0

Table 3: Results of the two scoring systems for the features based on the focus group data

Both scoring systems had matching results for the relevance of most features. Because of the addition of two additional rating categories, score 2 enabled a more diverse and detailed result of each feature. As stated above, features with a score of 75% (or 12 points) or higher have been considered as relevant for user satisfaction. When comparing the results of both scores, high consensus was reached for certain features. 9 features (see Table 4) received positive relevance scores and thus, were ranked highly in both scoring systems. The research group decided to keep them as relevant features for further testing in the second experimental phase.

Feature Number	Name
F 1	Response time
F 4	Privacy & security
F 5	Perceived credibility
F 6	Understandability
F 9	Ability to maintain themed discussion
F 10	Maxim of quantity
F 11	Ease of use (general)
F 12	Flexibility of linguistic input
F 15	Expectation setting

Table 4: The 9 features that received the highest scores by the participants of the focus group

The 5 factors (*F2, F14, F19, F20, F21*) that did not reach the 75% mark for both scores were considered to be removed from the list. Expert review decided to keep *F19 (Graceful responses in unexpected situations)* for further testing despite the low scores (see Table 3). It was named several times to be misunderstood by participants during the discussions.

F16 and *F18* had the largest difference in ranking between both scoring systems. Both features were discussed within the group of experts and to be found relevant for further testing (see Table 3).

The discussion of the results amongst the research group resulted in *Ease of use* (F11) and *Trust* (F3) to be removed from the preliminary list of features. These features are expected to be found present as an overall concept in every other feature.

Moreover, the features *Personality* (F20), *Enjoyment* (F21) and *Appropriate degrees of formality* (F8) were decided to be removed from the preliminary list but to be suggested to a future list of relevant features for chatbot designers.

Furthermore, F13, F7 and F17 were found to be ranked on the bottom of the second third. Therefore, the group of experts reconsidered these features. *Maxim of relation* (F7) and its meaning was found to be discussed often amongst the participants. Hence, this feature was added to the list of preliminary features for further research. *Visibility* (F13) was marked as irrelevant by end users, possibly because it's definition only applied to a certain type a chatbot – on a website. The research group agreed on keeping this feature for further evaluation during the

second experimental phase but making it more inclusive in order to fit every type of implemented information-retrieval chatbot. *Process tracking* (F17) was not considered further in this research. Multiple participants stated this feature to be unnecessary with appropriate and fast response times of chatbots. Moreover, the tasks for information-retrieval chatbots did not appear complex enough to consider process updates a relevant feature.

These quantitative results of the first experimental phase and the related decisions of the research group led to a preliminary list of 14 features for further investigation in the second experimental phase of this research (see Table 5). Prior feature descriptions for *F4, F7, F13* and *F19* are found to be unclear and result in misunderstandings. Their descriptions need to be rewritten in order to create a clear and understandable definition.

Feature Number	**Feature Name (Named Before)**	**Feature Description**
F1	Response time	Ability of the chatbot to respond timely to users' requests
F4	Perceived privacy & security (Privacy and security)	*Unclear*
F5	Perceived credibility	How correct and reliable the chatbot's output seems to be
F6	Understandability	Ability of the chatbot to communicate clearly and is easily understandable
F7	Relevance (Maxim of relation)	*Unclear*
F9	Ability to maintain 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 input

F10	Maxim of quantity	Ability of the chatbot to respond in an informative way without adding too much information
F12	Flexibility of linguistic input	How easily the chatbot understands the user's input, regardless of the phrasing
F13	Accessibility (Visibility)	*Unclear*
F14	Ease of starting a conversation	How easy it is to start interacting with the chatbot / to start typing
F15	Expectation setting	Make purpose clear, show user what it can and cannot do with chatbot, was taken from maxim of manners
F16	Reference to service	Ability of the chatbot to make references to the relevant service, for example, by providing links or automatically navigating to pages
F18	Recognition and facilitation of users' goals and intent	Ability of the chatbot to understand the goal and intention of the user and to help him accomplish these
F19	Handling unexpected situations (Graceful responses in unexpected situations)	*Unclear*

Table 5: Preliminary list of relevant features and descriptions after first experimental phase

Main findings for items.

Questionnaire items were generated for every feature. Results of the focus group showed items being ambiguous and thus, being matched with two or more features by the participants. A change in formulation at this point of the research would not guarantee the matching of an item with a single feature. Therefore, all previously generated items were kept for use in the questionnaire during the second experimental phase. Weak items were to be reassessed during the usability testing and expected to produce similar results.

Second Experimental Phase (Usability Testing)

Methods

Participants.

As for the pilot testing of the questionnaire, data of 59 participants was included. 62.71% of the participants were identifying as male (n=37) and 37.29% as female (n=22). The age ranged from 18 until 55 years (M=23.72, SD=4.88). All participants were recruited through convenience sampling. Moreover, 44.07% of the people were German (n=26), 5.08% Dutch (n=3) and 50.85% (n=30) from mostly non-European national backgrounds. The most prominent nationality of these was Indian (n=23). 32.2% (n=19) of the participants were currently enrolled in the University program of Psychology, while 66.1% (n=39) stated to belong to other working or study fields. It can be said that most other participants were following a technical university program (n=30).

Material.

The Participant Information Sheet (PIS) with informed consent (see Appendix A) was reused for the participants of the second phase. The questions of the Demographics questionnaire (see Appendix A) were embedded into the online survey software Qualtrics.

The experimental sessions were conducted on standard Laptops of the brands Asus and Acer with a Windows 8 and 10 OS. The screen size was either 13.3” or 15” inch. An English key board was used for all sessions. In order to record the computer screen and the participant’ speech, the Windows in-built game bar was used. Additionally, the experimenter took short notes of the main steps and occurring events per chatbot during the sessions.

The experimenter instructed the participant according to the script. This was originally developed and used for the research by Tariverdiyeva and Borsci (2019). A final version based on the original script was created through expert review of the research team (see Appendix B).

The students were asked to follow instructions provided online. In total, nine chatbots were tested for this research (see Appendix I). The two highest rated (ATO, Amtrak) and the two lowest rated (Toshiba, Inbenta) conversational agent services were chosen to be included from

the previous exploratory research of Tariverdiyeva and Borsci (2019) in order to ensure comparable and reliable results of the newly developed measurement tool.

The researcher then chose randomly one out of two conditions (A and B) which led to a randomized number of 5 chatbots and their belonging tasks being presented to the participant according to the survey flow (see Appendix B). The questionnaire was developed from the features and items tested in the previous phase of this research with 14 preliminary features (see Table 5) and the initial 42 items (see Appendix A). Moreover, it included the standard usability measurement scale UMUX-Lite (Lewis et al., 2013) with two items and a simple scale measuring task difficulty (0-100) with 1 item.

Procedure.

Before the arrival of the participant, the PIS was printed and the Qualtrics online questionnaire started. The researcher was randomly assigning the experimental condition, A or B. Additionally, the Windows game bar was started and set up.

The participant was welcomed by the researcher and received a short introduction, including the aim of the research itself and procedure of the experimental session. The participant was informed about the recording and consented to the experimental conditions found in the PIS. Now the participant was asked to open the Qualtrics questionnaire and follow the given instructions closely.

After the Demographics questionnaire, the online software led the student to the first task and chatbot. The chatbot was opened by copying the link into a new tab. The information indicated to be found was stated clearly in the task. The participant was able to stop the procedure of information retrieval when the information was found or the participant was not able to reach the goal after several attempts.

Afterwards the participant went back to the online questionnaire and started filling in the new developed tool UMT/C and UMUX-Lite (Lewis et al., 2013) to evaluate his or her experience with the just used chatbot. Moreover, the online software asked to fill-in a scale measuring the perceived task difficulty on a scale of 1-10.

The procedure was repeated for five randomly assigned chatbots per experimental session. During the time of using the chatbots, the attending researcher was available for

questions. In the end, the questionnaire indicated the end of the session. The experimenter then concluded the session and provided the participant with the option to contact the research team for further questions.

Data Analysis.

All procedures of the data analysis for the data set of the usability testing is conducted with the software program IBM SPSS Statistics 25. As a first step, all unfinished questionnaire data sets with a process of less than 100% were removed before the analysis.

Furthermore, Cronbach's Alpha was first calculated for the new tool in order to test the reliability of the overall questionnaire (Gliem & Gliem, 2003). UMUX-Lite scores were gathered for this research and scores calculated with the formula provided by Lewis et al. (2013). Thus, the results of chatbot evaluation can be compared to possible SUS results.

Questionnaire data of the new UMT/C was assessed by calculating mean scores of all 42 items per chatbot. Afterwards, results of the UMUX-Lite and the UMT/C could be compared and tested for possible correlation between the two measurements of user satisfaction.

Additionally, the mean scores of the UMT/C per chatbot were tested for possible correlation with the gathered data on perceived task difficulty per chatbot. It was expected for lower graded chatbots to be less helpful and thus, receive higher task difficulty scores. In order to test for influence of the population, the connection between the familiarity of participants with chatbots in general and the perceived task difficulty was estimated. Less familiarity should lead to a higher task difficulty perception.

Based on the previous study of Tariverdiyeva and Borsci (2019) and the results of the first experimental phase of this research, the questionnaire is presumed to measure several underlying factors that can be connected to different aspects of user satisfaction. It was unclear if the framework provided a complete and total set of features to investigate the factor loadings of the preliminary version of the questionnaire. Thus, an exploratory factor analysis (EFA) was applied to the data set, consisting of mean scores of each questionnaire item. The number of items was supposed to be reduced to a smaller set of possibly correlated variables – the underlying factors (Williams, Onsmann & Brown, 2010).

Despite the expectation to find certain factors that have been used to create the questionnaire, the correlation between them was still unknown. Items that showed more than one factor loading were removed. Despite that, at least one item was kept for every found factor.

In order to have external validation of the new measurement tool over the shorter 2-item scale UMUX-Lite (Lewis et al., 2013), the mean score of every factor was tested for correlation with the mean score provided by the UMUX-Lite.

The preliminary list of factors and their questionnaire items after conducting the EFA was used to estimate the internal reliability of each subscale with more than one questionnaire item.

Results

Demographic Analysis.

All participants of the second experimental phase filled in information about their previous interactions with chatbots. 38.98% participants rated their level of familiarity with chatbots to be moderately familiar (n=23). 55.93% of the participants (n=33) stated to have definitely used a chatbot before. 28.81% (n=17) have probably used one before and 15.24% (n=9) have not used one or are unsure. Next to that, it was asked how often chatbots were used in the past. The population on average had rarely used chatbots (n=37, M=4.58).

Questionnaire UMT/C.

The internal consistency of UMT/C was measured for the total number of 42 items and 9 chatbots. The range for Cronbach's α was estimated at a range of .885-.955 (M= 0.929) for the reliability of the questionnaire.

The end scores of UMT/C of every tested chatbot were calculated by combining the mean scores of every questionnaire item (see Table 6 and Figure 1). The Facebook messenger chatbot of Booking.com ranked the highest with an approximate score of 4.13. The lowest score with approximately 2.97 received the chatbot service "Emma" of USCIS (United States Citizenship and Immigration Services).

Tested Chatbot	Mean score	SD
Amtrak	3.6702	.51991
Toshiba	3.9215	.38476
ATO	3.9780	.50035
Booking	4.1267	.48102
Flowers	3.0411	.59836
Inbenta	3.2522	.75830
HSBC	3.6156	.62093
USCIS	2.9654	.67433
Absolut	3.2824	.60752

Table 6: The overall scores of chatbots calculated with the new measurement tool UMC/T

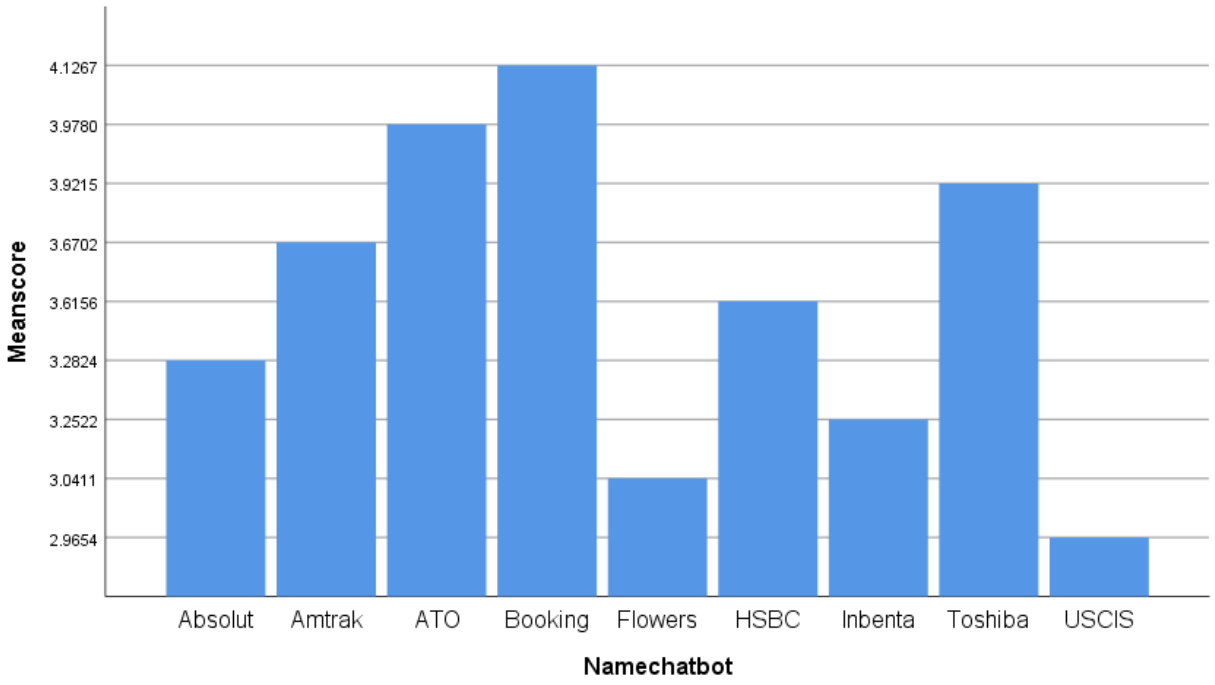


Figure 1: The tested chatbots and their estimated scores for usability measured with UMT/C

When testing for a correlation between the UMT/C score of every chatbot and the perceived task difficulty (see Table 7) for the end user, a clear positive and significant correlation of 0.7 or higher was found for the chatbots of ATO, HSBC and Inbenta. The other chatbot

services (Amtrak, Toshiba, Absolut, Booking, USCIS) showed moderate correlation scores with perceived task difficulty between 0.4 and 0.69. The chatbot of 1-800-Flowers was the only chatbot to not have a significant correlation with the scores of task difficulty at all.

Chatbot	**Pearson Correlation of UMT/C with Task Difficulty**	**Sig. (2-tailed)**
Amtrak	.647	.001
Toshiba	.415	.031
ATO	.706	.001
Booking	.444	.018
Flowers	.290	.127
HSBC	.832	.001
Inbenta	.758	.001
USCIS	.569	.001
Absolut	.668	.001

Table 7: Correlation between the UMT/C score and perceived task difficulty per chatbot

Participants of the usability test also stated their level of familiarity with chatbots or other conversational interfaces. A possible correlation of it with the perceived task difficulty was estimated. No statistically significant or clear correlation was found between these two measurements.

In order to assess the quality of the new questionnaire, it was tested for its correlation with the validated standard scale UMUX-Lite (Lewis et al., 2013) (see Table 8). The correlations of the UMT/C and UMUX-Lite scores were found to be statistically significant for every chatbot. All chatbot scores of UMT/C showed at least moderate correlation with the UMUX-Lite results. Two-thirds of the chatbot scores were estimated to have a positive correlation score of 0.79 or higher (on a scale of 0 to 1). The result of the chatbot service of 1-800-Flowers showed the lowest correlation with the UMUX-Lite score.

Chatbot	Pearson Correlation of UMT/C with UMUX-Lite	Sig. (2-tailed)
Amtrak	.844	.001
Toshiba	.824	.001
ATO	.791	.001
Booking	.823	.001
Flowers	.694	.001
Inbenta	.843	.001
HSBC	.856	.001
USCIS	.795	.001
Absolut	.829	.001

Table 8: The correlation between the new UMT/C scores and the UMUX-Lite results per chatbot

Exploratory Factor Analysis (EFA).

The mean score of the questionnaire items 1 to 42 was assessed for their factor loadings through EFA (see Table 9). Nine underlying factors were found for the UMT/C questionnaire. The orthogonally rotated matrix shows the items with moderate or high factor loading on *Factor 1-6*.

Q11 measures *Factor 9* moderately but shows a negative factor loading. *Q13* and *Q35* (in Table 9 underlined) load moderately onto an additional factor but are the highest-scoring items measuring *Factor 7* and *Factor 8*.

Every other item that loaded onto more than one factor was simply removed from the preliminary list for the questionnaire. In order to assess user satisfaction and the quality of user interaction with a chatbot, the chosen items have to avoid possible ambiguity of their results.

	Fa 1	Fa 2	Fa 3	Fa 4	Fa 5	Fa 6	Fa 7	Fa 8	Fa 9
Q1		.738							
Q2		.791							
Q3		.747							
Q4		.674							

Q5		.810						
Q6		.758						
Q8					.768			
Q11								-.685
Q13			.433			.582		
Q16	.817							
Q17	.635							
Q18	.721							
Q19			.817					
Q20			.820					
Q21			.805					
Q23	.770							
Q24	.769							
Q32				.585				
Q35	.353						.697	
Q37	.874							
Q38	.730							
Q39	.816							
Q40			.715					
Q41			.942					
Q42			.934					

Table 9: The factor loadings of each questionnaire item after factor analysis in a rotated matrix

Based on the results of the EFA, questionnaire items were matched to the 9 factors (see Table 10). The factors were renamed based on the content of the items.

Factor 3 (Perceived Privacy & Security) and *Factor 4* (Response Time) were measured as intended by the 3 generated items in the questionnaire. *Factor 5* to *Factor 9* were measured by one item each. They were belonging to the features *Handling unexpected situations, Expectation setting, Ability to maintain themed discussion, Understandability* and *Flexibility of Linguistic Input*.

Factor 1 included 8 items - 3 previously belonged to the feature *Reference to service*, 2 to

Recognition of User's Intent & Goal and 3 to *Perceived Credibility*. *Factor 2* consisted of 6 items in total. All 3 generated items of the features *Ease of starting the conversation* and *Accessibility*.

Factor	**New Name**	**Items**	**Represented Features**
Factor 1	Perceived credibility, Implementation & Understanding the User's Intent	Q16	F16 (Reference to service)
		Q17	F16
		Q18	F16
		Q23	F18 (Recognition of User's Intent & Goal)
		Q24	F18
		Q37	F5 (Perceived credibility)
		Q38	F5
		Q39	F5
Factor 2	Accessibility & Starting the conversation	Q1	F14 (Ease of starting the conversation)
		Q2	F14
		Q3	F14
		Q4	F13 (Accessibility)
		Q5	F13
		Q6	F13
Factor 3	Perceived Privacy & Security	Q19	F4 (Perceived Privacy & Security)
		Q20	F4
		Q21	F4
Factor 4	Response Time	Q40	F1 (Response Time)
		Q41	F1
		Q42	F1
Factor 5	Handling unexpected situations	Q32	F19 (Graceful Responses in Unexpected Situations)
Factor 6	Expectation setting	Q8	F15 (Expectation setting)
Factor 7	Ability to maintain themed discussion	Q13	F9 (Ability to maintain themed discussion)

Factor 8	Understandability	Q35	F6 (Understandability)
Factor 9	Flexibility of Linguistic Input	Q11	F12 (Flexibility of Linguistic Input)

Table 10: The new factors and the items that are found to be measuring them

In order to demonstrate the advantages of the UMT/C over the shorter UMUX-Lite tool (Lewis et al., 2013), the mean scores of the items of every factor were tested for their correlation with the mean UMUX-Lite scores (see Table 11).

Factor 1 shows a high significant correlation with the UMUX-Lite score. *Factor 2* and *Factor 7* still show moderate significant correlations. *Factor 3*, *Factor 4*, *Factor 6* and *Factor 8* correlate significantly but weakly with the 2-item scale.

Factor 5 and *Factor 9* show no statistically significant correlation with the UMUX-Lite results.

	Pearson Correlation of Factors with UMUX-Lite	**Significance Level (2-tailed)**
Factor 1	.713	.001
Factor 2	.435	.001
Factor 3	.345	.009
Factor 4	.378	.004
Factor 5	.095	.484
Factor 6	.383	.004
Factor 7	.501	.001
Factor 8	.384	.003
Factor 9	-.252	.061

Table 11: The correlation scores between the new factors and the UMUX-Lite scores

Cronbach's Alpha was estimated for each subscale with more than one item of the final questionnaire. The internal consistency of *Factor 1* to *Factor 4* was found to range between .889 and .927 (M=.904). The *Factors 5* to *9* are currently measured by only one questionnaire item.

Thus, the internal reliability of these subscales of the UMT/C could not be estimated for this research.

Discussion

The main goals of this research were to (a) *retest the relevant features that are found to influence user satisfaction for information-retrieval chatbots* and to (b) *test and possibly create a preliminary list of questionnaire items that can be used in a future quality measurement tool for chatbots – UMT/C*.

The initial part of this work enabled to reduce the list of 18 features taken from the previous research of Tariverdiyeva and Borsci (2019) to a list of 14 relevant key features. Based on expert review and the data gathered from the focus groups, main adjustments were made in terms of naming and formulating clear descriptions for these 14 key features. Also, it was found that the additional feature '*Expectation setting*' (F15), taken from the research of Luger and Sellen (2016), was considered to be relevant by end users and thus, it was added to the preliminary list. Therefore, the relevance of 13 out of 18 features previously found by Tariverdiyeva and Borsci (2019) was confirmed by this research.

However, four feature descriptions were found to be insufficient and unclear for possible end users. Thus, the research group recommends to reformulate the descriptions to avoid ambiguity and to test their quality in a future research.

A new version of the questionnaire UMT/C was generated in the second experimental phase with the data of the focus group. It showed a strong correlation with the standard usability scale UMUX-Lite (Lewis et al., 2013). The results of the new tool UMT/C can thus be considered to be valid and reliable when measuring the usability of a software.

UMT/C was used in the second experimental phase to assess the quality of user experience of nine chatbots. Four of the chatbots were taken over from the previous research of Tariverdiyeva and Borsci (2019) in order to evaluate the reliability of the new tool. As in the previous research, the chatbots of ATO and Amtrak received high scores from participants of the

usability testing. Also, the chatbot of Inbenta was expected to end in the lower ranks. Thus, these results measured with the new tool confirm the findings of the previous research. Surprisingly, the chatbot service of Toshiba was rated as one of the lowest in the previous research but was ranked in the third place when retested now. The discrepancy in perceived user satisfaction for this chatbot in particular should be assessed in future research.

The second research question was '*What questionnaire items should be included in the new measurement tool from an end user's perspective?'*. In order to answer this question, an Exploratory Factor Analysis (EFA) was applied to the data in the second experimental phase and allowed further insights into the underlying concepts of the new measurement tool UMT/C. Nine factors were found to be measured by in total 25 items. Thus, it can be said that the final 25 items belonging to the nine factors are forming a first preliminary questionnaire to reliably evaluate user satisfaction for information-retrieval chatbots.

In the nine factors captured by the new questionnaire, 12 out of 14 key features indicated by the first experimental phase were still included. However, '*Maxim of relevance' (F7)* and '*Maxim of quantity' (F10)* were not captured by the new questionnaire. Results of the focus group suggested that *F10* should be considered highly relevant and easy to understand for possible end users. Results for *F7* from the focus group were suggesting ambiguity because of low ratings of relevance and misunderstanding of the meaning from the participants of the focus groups. It also needs to be considered that both features were found to be highly related by end users of this research. Therefore, *F10* should be implemented into the questionnaire in order to capture the aspect of information quantity.

When measuring the correlation between each of the nine factors with the UMUX-Lite score, it was found that three factors (*Perceived credibility, Implementation & Understanding the User's Intent*; *Accessibility & Starting the conversation* and *Ability to maintain themed discussion*) correlate at least moderately with the standard usability scale. Thus, the results of the new measurement tool include the concepts measured by the standard 2-item usability scale. However, four factors only showed a weak correlation with the UMUX-Lite scale. *Perceived Privacy & Security, Understandability, Expectation setting* and *Response Time* seem to be additional features that extend simple usability measurement scales and cannot be captured by other standard tools.

Especially *Response Time* and *Understandability* can be seen as significant features of a

conversational software. Short response time and being able to understand the answers are most likely having a great impact on the motivational aspect for users. As found by Brandtzaeg and Følstad (2017), users are motivated to use chatbots when it is increasing their productivity. Fast and understandable responses thus are important features of software that communicates with the user and should be part of a usability measurement tool.

While the key feature *Perceived Privacy & Security* extents sole functionality of a software, it has to be considered when data is being shared online. Miyazaki and Fernandez (2001) suggest that privacy and security of data is user's number one concern when acting online. Chatbots specifically allow the user to share private information in a conversation. Naturally, the user has to have the impression that personal data will not be exposed in order to be satisfied with the service.

The study of Zamora (2017) confirms that meeting the expectations of users is a significant influence on the level of satisfaction. 50% of the participants of that research stated to not use a chatbot again because their prior expectations were not met. Therefore, *Expectation setting* is an influential feature that must be included in the new measurement tool.

The two other factors *Handling unexpected situations* and *Flexibility of Linguistic Input* could not be proven to be correlated with the UMUX-Lite results from the sample taken for this research. It is known that these features are captured by the new tool UMT/C but it needs to be investigated further if the same is true for other usability measurement tools.

Additionally, the concept of perceived task difficulty was assessed and found to be positively correlating with the results of the UMT/C in 88% of the cases. Thus, it is possible that the rating of usability is depending on how easily or how badly the chatbot guides the user to his or her goal. The better the usability of a chatbot, the easier a task will be perceived by the end user. On the other hand, the correlation could also be interpreted as the level of task difficulty having a direct effect on the perceived usability of the chatbot for an end user. Thus, the easier the task is perceived, the better the chatbot is rated. Czerwinski, Horvitz and Cutrell (2001) conducted a study on the relationship between perceived duration and success rate when rating the usability of a browser. It was found that the higher the failing rate for a task, the longer the time it took to conduct the task was overestimated. As a result, the time estimation could be used an implicit measurement of usability. Thus, it is most likely that the level of perceived task difficulty and the rate of success directly influenced the rating of the usability of a chatbot. This

needs to be considered when creating tasks for end users in future studies.

The level of familiarity of the end user with chatbots was investigated and found to not have an impact on the level of perceived task difficulty. This is contradicting findings that show that experience has a direct effect on perceived ease of use of a system (Hackbarth, Grover & Yi, 2003). The success rate of the user thus should have been directly linked to his or her expertise. Possibly no correlation was found in the present research because the participants of the usability testing were young adults. It can be assumed that, even though the experience with chatbots specifically was rated low, the general expertise with conversational software and technology was above average.

Limitations

The participants were asked to fulfill a simple task with a chatbot during the second experimental phase. The specific choice of task and the choice of words for each task might have influenced the rating of perceived task difficulty. Additionally, social dynamics and possible groupthink might have played a role during the first experimental phase. This could have led to a less accurate result for the focus group sessions. Because of limited time and resources, this research was conducted with a small number of participants. Moreover, the majority of participants was highly educated and following a university study which limits the possibility of generalization of the results. It should also be considered that it was attempted to include every concept related to measuring user satisfaction for chatbots but there is a possibility that not every feature was considered yet.

Recommendations

On the basis of this research and its findings, it is recommended to retest and confirm the results with a larger number of people and a more diverse pool of participants. Furthermore, the findings should be validated by assessing the new questionnaire with an additional number of tasks per chatbot. Also, the influence of the complexity and difficulty of the tasks needs to be taken into account.

Future research should focus on refining the list of questionnaire items to capture every

aspect of measuring user satisfaction. Moreover, a special focus could be to generate additional questionnaire items in order to estimate the reliability of the subscales of the UMT/C.

Additionally, the correlation between the factors extracted by Exploratory Factor Analysis needs to be addressed. Following studies are advised to retest the quality of the feature descriptions to avoid ambiguity in understanding.

While this research study was focusing on information-retrieval chatbots, future research could be extended to include other types of conversational agents. It needs to be assessed if the questionnaire items of UMT/C can be applied to reliably measure usability of other types of chatbots in the future.

Conclusion

Chatbot services are offered online but no scale has been developed to measure usability and user satisfaction reliably from an end user's perspective. A preliminary version of the new usability measurement tool - UMT/C - was created and compared to the standard scale UMUX-Lite (Lewis et al., 2013). It was proven to be comparable and reliable. Based on the Exploratory Factor Analysis nine underlying factors were found which are measured by 25 items. While the overall results of the UMT/C still correlated with the UMUX-Lite, the single factors measured additional features that were not captured yet by other standard tools. Thus, the new measurement tool is capturing usability and user satisfaction better, specifically for chatbots.

Future studies need to refine and finalize the list of questionnaire items used for the UMT/C. Also, the correlation between the nine factors needs to be assessed. In the future, this new tool will allow fast and accurate testing of the quality of interaction with chatbots from an end user's perspective in order to measure user satisfaction.

References

Brandtzaeg P.B. & Følstad A. (2017). *Why People Use Chatbots.* In: Kompatsiaris I. et al. (eds) Internet Science. INSCI 2017. Lecture Notes in Computer Science, vol 10673. Springer, Cham. https://doi.org/10.1007/978-3-319-70284-1_30

Chaves, A. P., & Gerosa, M. A. (2019). *How should my chatbot interact? A survey on human chatbot interaction design.* arXiv preprint arXiv:1904.02743.1.

Czerwinski, M., Horvitz, E. & Cutrell, E. (2001). Subjective Duration Assessment: An implicit Probe for Software Usability. In *Submission cover for IHM-HCI* (p. 167-170), Redmond, USA.

Dale, R. (2016). The return of chatbots. Natural Language Engineering. 22(5): 811–817. Cambridge University Press 2016. doi:10.1017/S1351324916000243

Gliem, J. & Gliem, R. (2003). Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type-Scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education,* Ohio State University, Columbus, USA.

Hackbarth, G., Grover, V. & Yi, M. (2003). Computer playfulness and anxiety: positive and negative mediators of the system experience effect on perceived ease of use. Information & Management, Volume 40, Issue 3, 2003, p. 221-232, doi: 10.1016/S0378 7206(02)00006-X.

International Organization for Standardization (2018). Ergonomics pf human-system interaction – Part 11: Usability: Definitions and concepts (ISO Standard No. 9241-11). Retrieved from https://www.iso.org/standard/63500.html

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018, June). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018* (pp. 895-906).

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 2099). New York, New York, USA: ACM Press. https://doi.org/10.1145/2470654.2481287

Luger, E. & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose, California, USA. doi:10.1145/2858036.2858288

McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-69365-1_3

Miyazaki, A. D.; Fernandez, A. (2001). Consumer Perceptions of Privacy and Security Risks for Online Shopping. *Journal of Consumer Affairs. **35** (1)*: 38–39.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents. Retrieved from https://arxiv.org/ftp/arxiv/papers/1704/1704.04579.pdf

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

Vasconcelos, M., Candello, H., Pinhanez, C., & dos Santos, T. (2017). Bottester: Testing Conversational Systems with Simulated Users. Doi:10.1145/3160504.3160584.

Williams, B., Brown, T., & Onsman, A. (2010). Exploratory factor analysis: A five-step guide for novices. Australasian Journal of Paramedicine, *8(3*). Retrieved from http://ro.ecu.edu.au/jephc/vol8/iss3/1

Zamora, J. (2017). I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. *Human Agent Interaction Conference 2017.* Bielefeld, Germany.

Appendix A

Participant Information Sheet and Informed consent

Participant Information Sheet

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

Principal investigator: Divyaa Balaji

Co-investigator: Dr Simone Borsci

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

Purpose of the study

This study aims to develop and validate a new measure for evaluating user satisfaction with chatbot interactions. One of the main tasks is to determine the factors that are the most important for measuring this construct. This will be done so through qualitative data gathered through focus groups using end-users. This data will be used to inform the items that will eventually make up the questionnaire. The questionnaire will then be administered in a usability testing paradigm for further validation.

Your role as participant

Note that your participation is entirely voluntary. Refusal or withdrawal will involve no penalty, now or in the future. If you wish to withdraw yourself from the study at any point of the session, please simply inform the responsible researcher.

Involvement in this study is not related to any risks of physical or mental kind for you as the participant.

Your participation in the focus group includes giving your opinion on different factors and items that are important in the usability testing of chatbots. You will be asked to evaluate certain factors and match items to the factors you think they are related to.

As for the second part of the research, you are asked to perform a usability test on several chatbots using the developed measurement tool. The experiment is including you to perform certain tasks in a chatbot when asked. Afterwards, you will have to fill in the questionnaire developed for usability testing of information-retrieval chatbots.

Personal data

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

Videotaping and Questionnaire

The focus group sessions will be videotaped so that the research team can use this information generated by the moderated group discussions to perform data analysis and acquire insight into the research question being studied. When performing the usability testing, each participant's questionnaire data will be anonymized and securely stored for our research team to analyse. Additionally, each participant will be videotaped while performing usability testing with each chatbot and will capture the participant's thoughts as they perform the tasks. These video recordings will enable the research team to retrieve valuable information about how users perceive and interact with chatbots.

All data will be made anonymous before stored and secured on a separate hard drive to which the research team and supervisor will have access during the research period while writing bachelor and master theses. When data evaluation is finished, the access will belong solely to the supervisor. The

UNIVERSITY OF TWENTE.

research has the potential to be published and therefore, the data will have a retention period of approximately 12 months, when it is expected to be published. During the retention period, only the supervisor will have access to it.

Ethical review of the study

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

Contact details

Principal Researcher
Divyaa Balaji
d.balaji@student.utwente.nl

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

Consent Form for Assessing user satisfaction with chatbot interactions

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	Yes	No
Taking part in the study		
I have read and understood the study information dated [DD/MM/YYYY], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study will involve either (a) a video-recorded focus group or (b) a video-recorded usability session.	○	○
I am aware that my face and voice will be recorded and that this data will be treated with discretion until destroyed.	○	○
Use of the information in the study		
I understand that information I provide will be used for data analysis while writing bachelor and master thesis and for potential publication.	○	○
I understand that personal information collected about me that can identify me, such as [e.g. my name or where I live], will not be shared beyond the study team.	○	○
I agree that my information can be quoted in research outputs	○	○
Consent to be Audio/video Recorded		
I agree to be audio/video recorded.	○	○
Future use and reuse of the information by others		
I give permission for the video data that I provide to be archived in the BMS Lab so it can be used for future research and learning.	○	○

Signatures

_____	_____	_____
Name of participant [printed]	Signature	Date

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

_____	_____	_____
Researcher name [printed]	Signature	Date

UNIVERSITY OF TWENTE.

Focus Group Script

[Introduction]

Hello everyone! Thank you for coming here today.

My name is [INSERT NAME] and I'll be the moderator for today's group discussion. Just to give you a brief overview, this study is about measuring user satisfaction when interacting with a chatbot. There isn't a measurement tool for this yet so we'd like to know what factors are involved when users such as yourselves evaluate a chatbot. If you choose to go ahead today, a group of you will give us your input on the factors involved in determining user satisfaction.

I would also like to introduce my co-moderator for today: [INSERT NAME]. She'll take notes and assist me during the session.

[Informed consent]

It is mentioned in the informed consent but there's one aspect I'd like to explain further. We'd like to video record this session for our Master and Bachelor research. We will only use the videos as sources of data to analyze for our projects and no one else apart from our research team will be able to see or use these videos. More information is available in the informed consent.

So before we begin, I'd like you to read, fill in and sign the informed consent form in front of you. If you have any questions about it while reading, please feel free to ask them. It's important that you understand everything before signing it.

[Introduction to chatbots]

As I mentioned earlier, this study is about interacting with chatbots. Have you guys interacted with a chatbot before?

For the benefit of those who haven't, a chatbot is a kind of software program running on artificial intelligence. They're expected to be able to simulate a human-like conversation, using natural language. Chatbots generally return a response based on either voice or text input from a user.

There are different kinds. You might have heard about ones like Apple's Siri, which are voice-activated virtual assistants. Today though, we'll be focusing on chatbots you can use to search for information online or information-retrieval chatbots. They're commonly found on websites to help customers but they can also be found on Facebook, for example.

Now that you have an idea of what a chatbot is, you may be able to recollect if you've used one before.

[Demographics]

So before we jump into the discussion, please fill out this short form for us about yourselves.

[Discussion guidelines]

We'd like to remind you of a few guidelines for this session.

First, everyone's opinion is valued and important for this topic. There is also no such thing as a right or wrong opinion.

Second, everyone should get the chance to talk without interruptions.

Third, this is a discussion and thus, you do not have to talk to me the whole time. It is perfectly fine to look and talk to each other directly.

Finally, we've planned for a 2-hour session but there will be breaks in between which you can use to get coffee or go to the toilet.

[Interactive demonstration]

To get you guys started, we're going to spend about 10 minutes testing a chatbot right now. If you haven't used one before, this is your chance to get familiar with them. If you have, then you can refresh your memory about them. So I'd like you to discuss and agree on what to ask the chatbot, essentially decide how to interact with it, and I will communicate with the chatbot. I will start the conversation and then you will take over from there.

<< Reflect on the experience we just had with the chatbot >>

<< What stood out to you? What did you like (or not like) about it? >>

<< Any questions or doubts about chatbots? >>

[Discuss factors]

Looking at research papers, we found many factors that researchers think are important for user satisfaction when interacting with chatbots to find information online. We now want your opinion on these factors.

(Give each individual the list of 11 factors)

<< Which ones do you consider important and/or relevant for interactions with such chatbots? On your list, mark the factors that you think are relevant. Think about why a factor is relevant to you or not. >>

<< First off, do you understand all of the factors? Are the explanations clear? If not, help us reword them to make them clearer. >>

<< Let's discuss some of these factors a little more. Which factors did you mark as irrelevant? Why? >>

<< Do you believe that there are any factors that we missed in this list? >>

(Repeat with the remaining 11 factors)

Break (5 minutes, when it seems necessary or let group decide if they want one)

[Discuss factors and items]

I hope that by now, all of you are familiar with all the factors presented in the list. We will now give you a list of items we generated that could potentially be included in the final questionnaire.

(Give each individual the list of items)

<< What we would like you to do is to try and match each item to the factor you think it represents. >>

<< While doing so, we would also like you to take a look at the items themselves – do you understand them? Do you think any of them should be reworded or otherwise changed/removed? If so, why? >>

<< Mark the ones you think could be reworded better by writing an R beside them. If you know exactly how you would rephrase the item or what is bothering you about it, you can write it down in the comments column beside the item. Mark the items you are happy with and would want to include in the questionnaire with a tick beside the item. Likewise, mark the items that you would remove with a cross beside the item. This could be because the item just doesn't make sense or because it's redundant, for example.

Remember: (1) there are no right or wrong answers for this exercise – it's about your opinion so sort them according to your intuition, (2) several items can be matched to one factor and (3) not all items need to be matched to a factor

<< Are there any questions? >>

[End]

Thank you all for your participation and nice discussion today. You were really productive. Are there any questions? If you have questions later, you can still contact us via email.

Demographics Questionnaire

Where applicable, please circle your chosen response. If not, fill in your response manually.

Age _____

Gender M / F

Nationality _____

Field of study _____

Have you used a chatbot before? Yes / No

If yes, then answer the two questions below.

How often do you use chatbots?

1	2	3	4	5
Never	**Rarely**	**Sometimes**	**Often**	**Always**

How would you rate your previous experiences with chatbots?

1	2	3	4	5
Very poor	**Poor**	**Fair**	**Good**	**Excellent**

List of features and descriptions

Session: Participant ID:

No.	Factor	Description	Relevant?	Why or why not?
1	Response time	Ability of the chatbot to respond timely to users' requests		
2	Engage in on-the-fly problem solving	Ability of the chatbot to solve problems instantly on the spot		
3	Trust (general)	Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage		
4	Privacy & security	Ability of the chatbot to protect the user's privacy		
5	Perceived credibility	How correct and reliable the chatbot's output seems to be		
6	Understandability	Ability of the chatbot to communicate clearly and is easily understandable		
7	Maxim of relation	Ability of the chatbot to provide the relevant and appropriate contribution to peoples needs at each stage		
8	Appropriate language style	Ability of the chatbot to use appropriate language style for the context		
9	Ability to maintain 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 input		
10	Maxim of quantity	Ability of the chatbot to respond in an informative way without adding too much information		
11	Ease of use (general)	How easy it is to interact with the chatbot		

Session: Participant ID:

No.	Factor	Description	Relevant?	Why or why not?
12	Flexibility of linguistic input	How easily the chatbot understands the user's input, regardless of the phrasing		
13	Visibility (website only)	How easy it is to locate and spot the chatbot on the website		
14	Ease of starting a conversation	How easy it is to start interacting with the chatbot / to start typing		
15	Expectation setting	Make purpose clear, show user what it can and cannot do with chatbot, was taken from maxim of manners		
16	Reference to service	Ability of the chatbot to make references to the relevant service, for example, by providing links or automatically navigating to pages.		
17	Process tracking	Ability of the chatbot to inform and update users about the status of their task in progress		
18	Recognition and facilitation of user's goal and intent	Ability of the chatbot to understand the goal and intention of the user and to help him accomplish these		
19	Graceful responses in unexpected situations	Ability of the chatbots to gracefully handle unexpected input, communication mismatch and broken line of conversation		
20	Personality	The chatbot appears to have a (human-like) personality		
21	Enjoyment	How enjoyable the interaction with the chatbot appears to be to the user		

List of questionnaire items

Q1: It was clear how to start a conversation with the chatbot.

Q2: It was easy for me to understand how to start the interaction with the chatbot.

Q3: I find it easy to start a conversation with the chatbot.

Q4: The chatbot was easy to access.

Q5: The chatbot function was easily detectable.

Q6 It was easy to find the chatbot.

Q7 Communicating with the chatbot was clear.

Q8 I was immediately made aware of what information the chatbot can give me.

Q9 It is clear to me early on about what the chatbot can do.

Q10 I had to rephrase my input multiple times for the chatbot to be able to help me.

Q11 I had to pay special attention regarding my phrasing when communicating with the chatbot.

Q12 It was easy to tell the chatbot what I would like it to do.

Q13 The interaction with the chatbot felt like an ongoing conversation.

Q14 The chatbot was able to keep track of context.

Q15 The chatbot maintained relevant conversation.

Q16 The chatbot guided me to the relevant service.

Q17 The chatbot is using hyperlinks to guide me to my goal.

Q18 The chatbot was able to make references to the website or service when appropriate.

Q19 The interaction with the chatbot felt secure in terms of privacy.

Q20 I believe the chatbot informs me of any possible privacy issues.

Q21 I believe that this chatbot maintains my privacy.

Q22 I felt that my intentions were understood by the chatbot.

Q23 The chatbot was able to guide me to my goal.

Q24 I find that the chatbot understands what I want and helps me achieve my goal.

Q25 The chatbot gave relevant information during the whole conversation

Q26 The chatbot is good at providing me with a helpful response at any point of the process.

Q27 The chatbot provided relevant information as and when I needed it.

Q28 The amount of received information was neither too much nor too less

Q29 The chatbot gives me the appropriate amount of information

Q30 The chatbot only gives me the information I need

Q31 The chatbot could handle situations in which the line of conversation was not clear

Q32 The chatbot explained gracefully when it could not help me

Q33 When the chatbot encountered a problem, it responded appropriately

Q34 I found the chatbot's responses clear.

Q35 The chatbot only states understandable answers.

Q36 The chatbot's responses were easy to understand.

Q37 I feel like the chatbot's responses were accurate.

Q38 I believe that the chatbot only states reliable information.

Q39 It appeared that the chatbot provided accurate and reliable information.

Q40 The time of the response was reasonable.

Q41 My waiting time for a response from the chatbot was short.

Q42 The chatbot is quick to respond.

Appendix B

Usability Testing Script

<<For researcher only: enter participant code and condition>>

Welcome to our study. We appreciate you helping us out today! We are in the process of developing a measurement tool to assess user satisfaction with information-retrieval chatbots. Today, you will be testing some chatbots and providing us with your feedback by responding to questionnaires. You will be presented with five chatbots, each with an associated task to do. After using each chatbot, you will have a few questionnaires to respond to. They will be presented to you through an online survey software. The session is expected to last for no more than 1.5 hours.

Remember that we will be recording you and the screen for data analysis purposes. If you are not okay with this, please let us know. There are more details in the informed consent which you must read and sign before proceeding further.

<<Give participant informed consent form>>

First, please fill in the demographic questionnaire.

You will now begin testing chatbots. Each provided task is a short realistic scenario – you, as the participant, should try your best to imagine yourself in those situations i.e. imagine that you're looking for that information for the first time. If you do not understand the situation or task, let me know. Once you feel like you have achieved the task, or if you feel that the task is not achievable, please let me know. You can then move onto the relevant questionnaires. I would like to emphasise that there is no wrong or right answer in this test. Your behaviour and responses will help us understand how users use and think about 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 at any point.

<<Start recording the screen>>

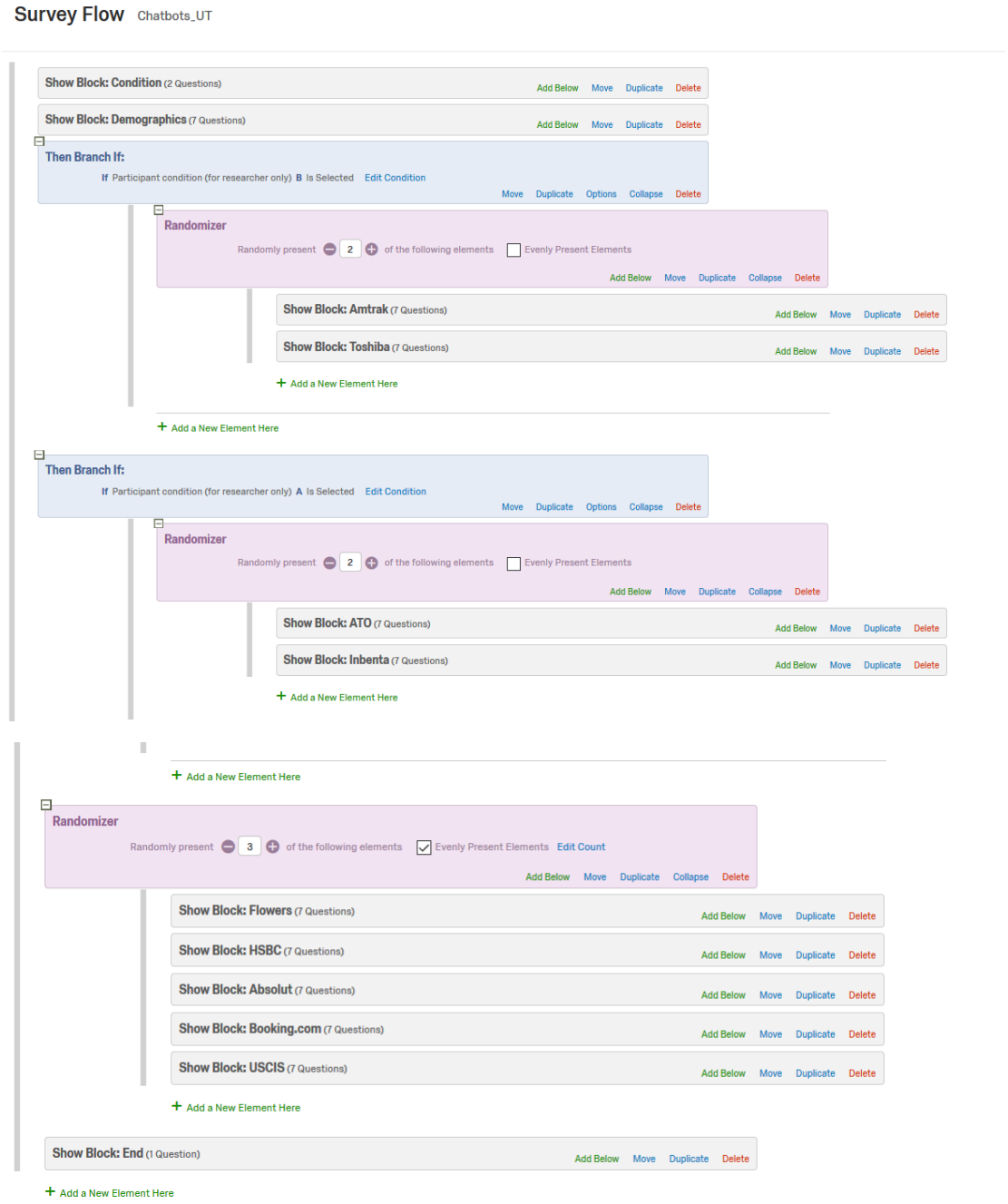
<<After finishing the questionnaire>>

Thank you for your participation.
Do you have any questions? Otherwise feel free to contact us via the given email address.

Have a nice day!

Qualtrics Questionnaire Flow

As visible in the picture, the participant was asked to answer questions about his or her demographics and prior experience with chatbots before the first interaction task. Afterwards, the flow continued depending on the chosen condition A or B with a randomized set of chatbots and their tasks.



Preliminary Universal Measurement Tool -UMT/C

Part 1 (Demographics and Prior experience with chatbots)

Chatbots_UT

Start of Block: Condition

Q87 Participant ID

_____

Q13 Participant condition (for researcher only)

- ○ A (1)
- ○ B (2)

End of Block: Condition

Start of Block: Demographics

Gender

▼ Male (1) ... Prefer not to say (3)

Age

_____

Nationality

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

Study Field of study

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

Familiarity

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

Prior_Usage

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

Display This Question:

If = Definitely yes

Or = Probably

Or = Unsure

How_often

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

Part 2 (Questionnaire items of Preliminary UMT/C to assess usability of chatbot)

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

It is clear to me early on about what the chatbot can do. (9)	○	○	○	○	○
I had to rephrase my input multiple times for the chatbot to be able to help me. (10)	○	○	○	○	○
I had to pay special attention regarding my phrasing when communicating with the chatbot. (11)	○	○	○	○	○
It was easy to tell the chatbot what I would like it to do. (12)	○	○	○	○	○
The interaction with the chatbot felt like an ongoing conversation. (13)	○	○	○	○	○
The chatbot was able to keep track of context. (14)	○	○	○	○	○
The chatbot maintained relevant conversation. (15)	○	○	○	○	○
The chatbot guided me to the relevant service. (16)	○	○	○	○	○

The chatbot is using hyperlinks to guide me to my goal. (17)	○	○	○	○	○
The chatbot was able to make references to the website or service when appropriate. (18)	○	○	○	○	○
The interaction with the chatbot felt secure in terms of privacy. (19)	○	○	○	○	○
I believe the chatbot informs me of any possible privacy issues. (20)	○	○	○	○	○
I believe that this chatbot maintains my privacy. (21)	○	○	○	○	○
I felt that my intentions were understood by the chatbot. (22)	○	○	○	○	○
The chatbot was able to guide me to my goal. (23)	○	○	○	○	○
I find that the chatbot understands what I want and helps me achieve my goal. (24)	○	○	○	○	○

The chatbot gave relevant information during the whole conversation (25)	○	○	○	○	○
The chatbot is good at providing me with a helpful response at any point of the process. (26)	○	○	○	○	○
The chatbot provided relevant information as and when I needed it. (27)	○	○	○	○	○
The amount of received information was neither too much nor too less (28)	○	○	○	○	○
The chatbot gives me the appropriate amount of information (29)	○	○	○	○	○
The chatbot only gives me the information I need (30)	○	○	○	○	○
The chatbot could handle situations in which the line of conversation was not clear (31)	○	○	○	○	○

The chatbot explained gracefully when it could not help me (32)	○	○	○	○	○
When the chatbot encountered a problem, it responded appropriately (33)	○	○	○	○	○
I found the chatbot's responses clear. (34)	○	○	○	○	○
The chatbot only states understandable answers. (35)	○	○	○	○	○
The chatbot's responses were easy to understand. (36)	○	○	○	○	○
I feel like the chatbot's responses were accurate. (37)	○	○	○	○	○
I believe that the chatbot only states reliable information. (38)	○	○	○	○	○
It appeared that the chatbot provided accurate and reliable information. (39)	○	○	○	○	○
The time of the response was reasonable. (40)	○	○	○	○	○

My waiting time for a response from the chatbot was short. (41)	○	○	○	○	○
The chatbot is quick to respond. (42)	○	○	○	○	○

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

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

List of tested chatbots

- Toshiba
  (http://www.toshiba.co.uk/generic/yoko-home/)
- Amtrak
  (https://www.amtrak.com/home)
- USCIS
  (http://www.uscis.gov/emma)
- HSBC UK
  (https://www.hsbc.co.uk/)
- Absolut Vodka
  (https://www.absolut.com/en/)
- Inbenta
  (http://www.inbenta.com/en/)
- Booking.com
  (https://www.facebook.com/messages/t/131840030178250)
- ATO
  (http://www.ato.gov.au/)
- 1-800 Flowers
  (https://www.facebook.com/messages/t/1800FlowersAssistant)

Tasks for the tested chatbots

Amtrak_Task
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.

Toshiba_Task
You have Toshiba laptop of Satellite family and you are using Windows 7 operating system on your laptop. You want to partition your hard drive because it will make it easier to organize your video & audio libraries

ATO_Task
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.

Inbenta_Task
You have an interview with Inbenta in a few days and you want to use Inbenta's chatbot to find out the address of Inbenta's Mexico office.

Flowers_Task
It is your 1st anniversary with your significant other but they are back in the Netherlands and you are on a business trip in France and you would like to send them blue flowers (it's their favourite colour). Remember that you have a budget of 40 dollars. You want to use the 1-800-Flowers Assistant chatbot to look at your options.

HSBC_Task
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.

Absolut_Task
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.

Booking_Task
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

USCIS_Task
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.

Appendix C

SPSS Syntax

Parts of the syntax is shown solely for the chatbot of ATO. In order to calculate the results for the other tested chatbots, the variable name was exchanged.

Computing the score for every chatbot with the preliminary UMT/C scale

```
COMPUTE ATOmean=MEAN(ATO_USQ_2,ATO_USQ_1,ATO_USQ_3,ATO_USQ_4,ATO_USQ_5,
    ATO_USQ_6,ATO_USQ_7,ATO_USQ_8,ATO_USQ_9,ATO_USQ_10,ATO_USQ_11,
    ATO_USQ_12,ATO_USQ_13,ATO_USQ_14,ATO_USQ_15,ATO_USQ_16,ATO_USQ_17,
    ATO_USQ_18,ATO_USQ_19,ATO_USQ_20,ATO_USQ_21,ATO_USQ_22,ATO_USQ_23,
    ATO_USQ_24,ATO_USQ_25,ATO_USQ_26,ATO_USQ_27,ATO_USQ_28,ATO_USQ_29,
    ATO_USQ_30,ATO_USQ_31,ATO_USQ_32,ATO_USQ_33,ATO_USQ_34,ATO_USQ_35,
    ATO_USQ_36,ATO_USQ_37,ATO_USQ_38,ATO_USQ_39,ATO_USQ_40,ATO_USQ_41,
    ATO_USQ_42).
EXECUTE.
```

Computing the score for every chatbot with the formula of UMUX-Lite

```
DATASET ACTIVATE DataSet1.
COMPUTE ATOUMUX=(MEAN(ATO_UMUX_1)+MEAN(ATO_UMUX_2)-2)/ 12*100.
EXECUTE.
```

Computing the correlation between the UMT/C and UMUX-Lite results

```
CORRELATIONS
  /VARIABLES=ATOmean ATOUMUX
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Computing the correlation between the UMT/C scores and Task Difficulty

```
CORRELATIONS
  /VARIABLES=ATOmean ATO_TD_1
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Computing the correlation of Familiarity with chatbots and Task Difficulty

```
CORRELATIONS
  /VARIABLES=Familiarity_1 ATO_TD_1
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Computing Cronbach's Alpha for the questionnaire items of the tested chatbots

```
RELIABILITY
  /VARIABLES=ATO_USQ_1 ATO_USQ_2 ATO_USQ_3 ATO_USQ_4 ATO_USQ_5 ATO_USQ_6
ATO_USQ_7 ATO_USQ_8 ATO_USQ_9 ATO_USQ_10 ATO_USQ_11 ATO_USQ_12
ATO_USQ_13 ATO_USQ_14 ATO_USQ_15 ATO_USQ_16 ATO_USQ_17 ATO_USQ_18
ATO_USQ_19 ATO_USQ_20 ATO_USQ_21 ATO_USQ_22 ATO_USQ_23 ATO_USQ_24
ATO_USQ_25 ATO_USQ_26 ATO_USQ_27 ATO_USQ_28 ATO_USQ_29 ATO_USQ_30
ATO_USQ_31 ATO_USQ_32 ATO_USQ_33 ATO_USQ_34 ATO_USQ_35 ATO_USQ_36
ATO_USQ_37 ATO_USQ_38 ATO_USQ_39 ATO_USQ_40 ATO_USQ_41 ATO_USQ_42
  /SCALE('.') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE
  /SUMMARY=TOTAL.
```

Creating a new variable that captures the mean of every questionnaire item for all chatbots

This procedure needs to be repeated for questionnaire items 2 to 42.

```
COMPUTE
USQ_1=MEAN(Amtrak_USQ_1,Toshiba_USQ_1,ATO_USQ_1,Inbenta_USQ_1,Flowers_USQ_1,Book
ing_USQ_1,
    HSBC_USQ_1,USCIS_USQ_1,Absolut_USQ_1).
EXECUTE.
```

Exploratory Factor Analysis (EFA)

```
FACTOR
  /VARIABLES USQ_1 USQ_2 USQ_3 USQ_4 USQ_5 USQ_6 USQ_7 USQ_8 USQ_9 USQ_10
USQ_11 USQ_12 USQ_13 USQ_14 USQ_15 USQ_16 USQ_18 USQ_17 USQ_19 USQ_20 USQ_21
USQ_22 USQ_23 USQ_24 USQ_25 USQ_26 USQ_27 USQ_29 USQ_30 USQ_31 USQ_32 USQ_33
USQ_34 USQ_35 USQ_36 USQ_37 USQ_38 USQ_39 USQ_40 USQ_41 USQ_42 USQ_28
  /MISSING PAIRWISE
  /ANALYSIS USQ_1 USQ_2 USQ_3 USQ_4 USQ_5 USQ_6 USQ_7 USQ_8 USQ_9 USQ_10 USQ_11
USQ_12 USQ_13 USQ_14 USQ_15 USQ_16 USQ_18 USQ_17 USQ_19 USQ_20 USQ_21 USQ_22
USQ_23 USQ_24 USQ_25 USQ_26 USQ_27 USQ_29 USQ_30 USQ_31 USQ_32 USQ_33 USQ_34
USQ_35 USQ_36 USQ_37 USQ_38 USQ_39 USQ_40 USQ_41 USQ_42 USQ_28
  /PRINT INITIAL CORRELATION KMO AIC EXTRACTION ROTATION
  /FORMAT BLANK(.3)
  /CRITERIA MINEIGEN(1) ITERATE(100)
  /EXTRACTION PAF
  /CRITERIA ITERATE(100)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.
```

Computing new variables for the new found Factors with the belonging items

```
DATASET ACTIVATE DataSet1.
COMPUTE Factor1=MEAN(USQ_16,USQ_18,USQ_17,USQ_23,USQ_24,USQ_37,USQ_38,USQ_39).
EXECUTE.

COMPUTE Factor2=MEAN(USQ_1,USQ_2,USQ_3,USQ_4,USQ_5,USQ_6).
EXECUTE.
```

```
COMPUTE Factor3=MEAN(USQ_19,USQ_20,USQ_21).
EXECUTE.

COMPUTE Factor4=MEAN(USQ_40,USQ_41,USQ_42).
EXECUTE.

COMPUTE Factor5=USQ_32.
EXECUTE.

COMPUTE Factor6=USQ_8.
EXECUTE.

COMPUTE Factor7=USQ_13.
EXECUTE.

COMPUTE Factor8=USQ_35.
EXECUTE.

COMPUTE Factor9=USQ_11.
EXECUTE.
```

Computing the correlation between the nine Factors and the UMUX-Lite Scores

```
COMPUTE UMUXAll=MEAN (FlowersUMUX, BookingUMUX, AbsolutUMUX, USCISUMUX,
HSBCUMUX, InbentaUMUX, ATOUMUX, ToshibaUMUX, AmtrakUMUX).
EXECUTE.
CORRELATIONS
  /VARIABLES=UMUXAll Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8 Factor9
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Computing the new score for Cronbach's Alpha after EFA

This procedure needs to be repeated for every subscale with more than one item.

```
RELIABILITY
  /VARIABLES= USQ_16 USQ_18 USQ_17 USQ_23 USQ_24 USQ_37 USQ_38 USQ_39
  /SCALE('.') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE CORR
  /SUMMARY=TOTAL.
```