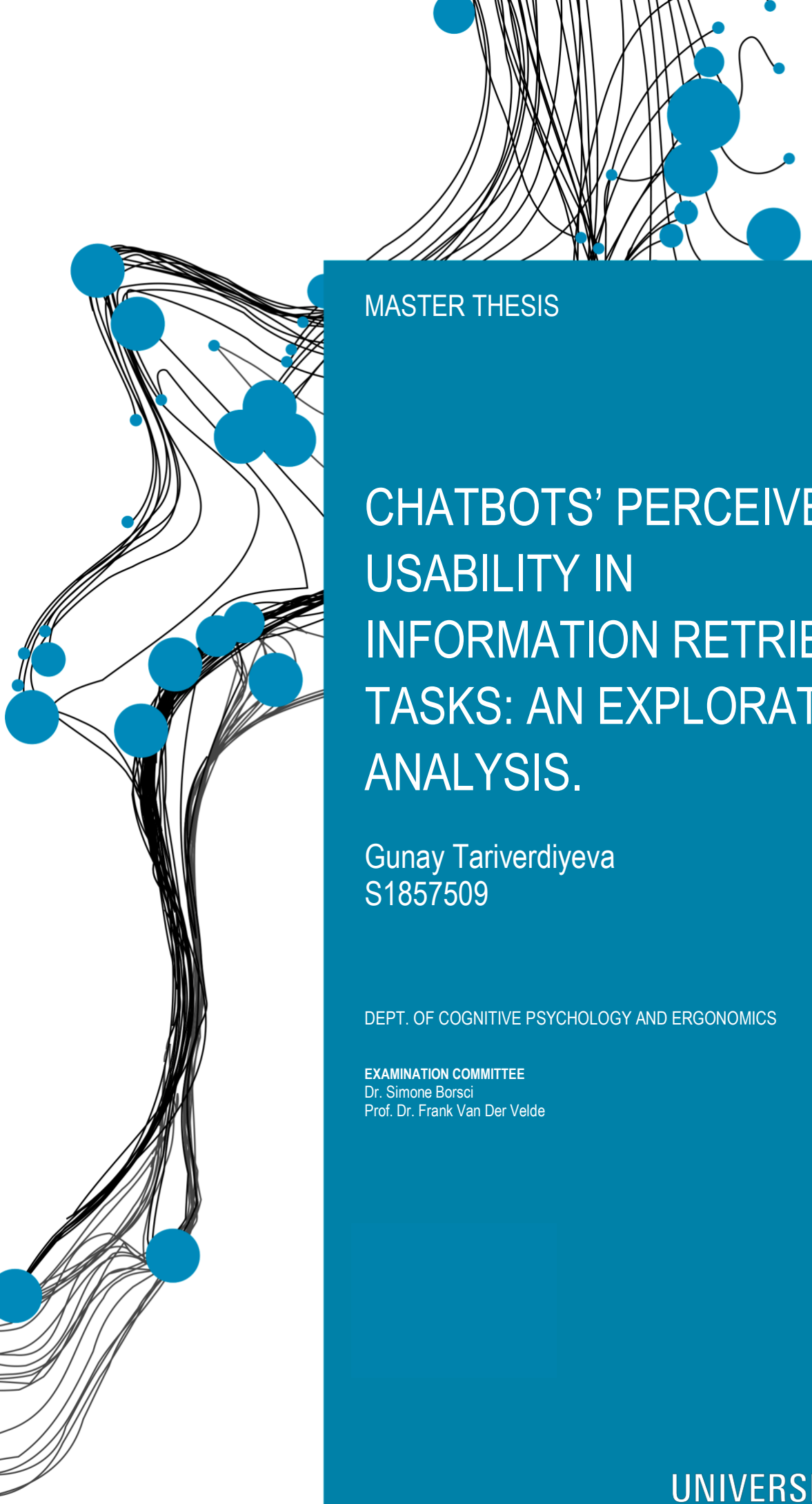


MASTER THESIS

CHATBOTS' PERCEIVED USABILITY IN INFORMATION RETRIEVAL TASKS: AN EXPLORATORY ANALYSIS.

Gunay Tariverdiyeva
S1857509

DEPT. OF COGNITIVE PSYCHOLOGY AND ERGONOMICS

EXAMINATION COMMITTEE
Dr. Simone Borsci
Prof. Dr. Frank Van Der Velde

UNIVERSITY OF TWENTE.

Chatbots' Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis.

Master Thesis

Gunay Tariverdiyeva (s185759)
Dept. of Cognitive Psychology and Ergonomics
University of Twente
g.tariverdiyeva@student.utwente.nl

Examination Committee:
Dr. Simone Borsci
Prof. Dr. Frank Van Der Velde

Date: 25.01.2019

Abstract

Despite many benefits, a substantial number of chatbots struggle to satisfy users. To understand what factors influence interaction with a chatbot and how to assess those factors, the present study aimed to; (a) explore the factors that are essential for user satisfaction, (b) investigate UMUX-Lite scale as a usability measure of a chatbot, and to (c) propose a design approach for a potential usability questionnaire. The research consisted of three phases: a systematic literature review, an online survey, and an interaction test. A comprehensive literature review identified 27 factors that influence interaction with a chatbot. Further, the initial list was filtered via the online survey in which 11 experts and 9 end-users participated. Next, the interaction test with 15 participants was conducted to get the perspective of users who just had experience with chatbots and to assess UMUX-Lite as a usability measure for chatbots. The online survey and interaction test distinguished 18 factors as important for satisfactory chatbot interactions. Furthermore, the current study found that UMUX-Lite is reasonably sensitive and a reliable measure of usability. However, as UMUX-lite does not include all the aspects important for measuring the perceived usability of the chatbot there is a concern about the extent of the sensitivity and validity of the scale. These findings suggest a need for a post-test questionnaire that will be able to capture more aspects of interaction with chatbots. Therefore, the present study recommends the development of a new questionnaire that will incorporate UMUX-Lite and proceed further on the basis of the key factors established in this study.

Keywords: Chatbot; conversational agents; usability; satisfaction; UMUX-Lite

Table of Contents

1. INTRODUCTION

Communication between humans through natural language is fascinating. It is not surprising that the way humans communicate with each other and perceive this communication has been studied for centuries. For example, by now it is well established that the style in which information is expressed can influence how we perceive information that is being communicated (Tubbs, 2013). However, what happens when our conversation partners are conversational agents (CAs)—bots developed to mimic human interaction? How do we perceive communication style that is so similar to the natural language we use, yet different? In contrast to humans, conversing in natural language is not an innate capability of robots—it is something that is programmed into them through complex programming techniques. Even though it is possible to track patterns in a conversation and create conversation schemas (McTear, Callejas, Barres, 2016) in order to teach machines how to chat with humans, it is often not enough to withstand the volatile nature of the natural language. Nevertheless, the "natural" aspect of the language— the possibility of the users to interact with technology in natural language—is what makes conversational agents so appealing. Use of the natural language as a mode of interaction can make technology much more accessible and effortless (Gnewuch, Morana, & Maedche, 2018). This is why studying the ways humans perceive interaction with conversational agents is important.

1.1 Conversational Agents

Conversational agents can be described as a "software that accepts natural language as an input and generates natural language as an output, engaging in a conversation with the user" (Griol, Carbó, & Molina, 2013). An interaction between the user and the agent happens through the conversational interface (CI). A conversational interface provides front-end to the CA and enables a user to interact with software by using speech, text, touch, and various other input and output modes (McTear, 2017). McTear, Callejas and Barres (2016) divided CIs into categories based on the communities that historically worked on CIs: spoken dialogue systems (SDSs), voice user interface (VUIs), embodied conversational agents (ECA), and chatbots. These communities often worked independently of each other and the types of CIs they developed differed in their goals and methods. Chatbots are CAs that simulate a conversation in natural language via text input and automatic text output (McTear, Callejas, & Barres, 2016). Originally chatbots were developed to try and fool users to believe that they are talking to another human. They were designed to maintain a small-talk and used a stimulus-response approach where users input is matched against a large set of stored patterns to generate a response (McTear et al.,

2016). Consequently, such an approach made chatbots more reliant on the user's input and less likely to initiate conversation.

Demand on chatbots is growing (Nguyen, 2017)and due to this demand, areas of application are expanding as well (e.g. education, information retrieval, customer service, e-commerce, health, finance). This creates new requirements that push chatbots to broaden their boundaries further and incorporate features from other types of conversational interfaces. Currently, many chatbots deploy more complex techniques, more initiative, are embodied (avatar, talking animation), use speech output that are characteristics to other CA communities.

1.2 The Rise of Chatbots

Dialogue-based systems have been around for a relatively long time, and programmers started to work on advancing speech-enabled interactive systems from the late-'80s (McTear et al., 2016). However, the rise and tipping point of chatbot development can be observed from 2016 (McTear, 2017). Only within launch year of bot platform on Facebook Messenger, the number of developed bots surpassed 100,000 and currently marks over 300,000 (Johnson, 2018). Rapid development in this area has been influenced by several points (a) advances in artificial intelligence (AI), (b) the availability of big data, (c) increased connectivity of devices and cloud-based resources, (d) advances in Speech Recognition and Natural Language Processing (McTear, 2017). Furthermore, many major tech companies opened their platforms to CAs and started investing heavily into technologies like artificial intelligence, deep learning, natural language processing, that are essential for building a system that can interact with the user in natural language (McTear, 2017). Apart from the technological breakthroughs and investment of big tech companies', interest to the CAs was also sparked by the shift in user behaviour towards messaging. In 2015 four top messaging apps exceeded four top social media networks with respect to global monthly active users (Nguyen, 2017).

Moreover, ease of implementation caused by technological developments and attention shift towards messaging platforms created a chance for the companies to reach their users in a new, more efficient, cost-effective and direct way (Toplin, 2017). WeChat, one of the most used apps in the Asian market, opened its platform for bots in 2013, and since then chatbots became one of the favourite ways for Chinese businesses to decrease workload that falls on the customer interaction (Van Eeuwen, 2017). In China WeChat enables its users to transfer money, order food, order a taxi, book a flight, all within its native app (Van Eeuwen, 2017). Even though adoption of chatbots is slower in the other parts of the world, there are still numerous cases

where chatbots succeeded and brought significant benefits both to the users and organisations around the world, for example:

- Sephora saw 11% increase in their makeover appointments (Kojouharov, 2018).
- More than 70% of orders that 1–800-Flowers gets over Facebook Messenger chatbot are from the new customers (Kojouharov, 2018).
- KLM raised customer interactions by 40% and helped 15% of customers to get their boarding passes (Kojouharov, 2018).
- Swedbank's chatbot, engages in 40,000 conversations in a month and solves 81% of them (Kojouharov, 2018).
- JPMorgan Chase saved over 360,000 hours of manpower in less than a year (Kojouharov, 2018).

1.3 Study Aims and Outline

Despite many success stories and hypothesised benefits, there is a substantial number of chatbots that struggle to deliver on their promise and disappear from the web (Araujo, 2018; Gnewuch et al., 2018). This raises a question—why do some chatbots succeed and others fail? The growth and failures CIs have experienced last five years opened an opportunity for new discussions and created further questions to answer. Currently, researchers are only catching up on the new challenges (Araujo, 2018). Moreover, an ambition to integrate CIs into the daily lives of the users has introduced novel challenges. Nevertheless, work on CAs promises to change the interaction of humans with technology as we know it today (Følstad & Brandtzaeg, 2017).

One question that needs more exploration is perceived usability of chatbots and its assessment. At present, there is a lack of literature on the tools that can be used to evaluate perceived usability of chatbot from user's perspective, that is practical and can be directly applied to the development process. For reasons beyond the scope of this thesis, there exists a communication gap between practitioners and academia and consequently, there is a lack of cohesive information on what factors are essential to designing a successful chatbot. Current academic literature is mostly focused either on the architecture and assessment of chatbot architecture or on very narrow aspects of the interaction between user and chatbot (e.g., Chakrabarti & Luger (2015); Huang, Li, Lin, & Yang (2015); Meira & Canuto (2015); Peeters (2016)). Such studies are not sufficient enough to inform and guide the developers throughout

design process. There is a need for studies that give a more extensive overview and practical tools that can be applied by designers on a daily basis.

RQ 1. What are the key factors that affect perceived usability in interaction with chatbots?

Furthermore, the literature review conducted in the present study did not identify a questionnaire or instrument specialised on evaluating perceived usability in interaction with chatbots that could be directly applied to the design process. Although there are generic tools to assess the usability of the system, thus fur not all these tools have been tested and shown to be valid and reliable in measuring the usability of chatbots. Therefore, it would be neglectful to assume that these tools are reliable and valid measures of usability in chatbots without testing. Further, the existence of standardised tools for usability assessment does not imply that there is no room for new methods and tools. New methods and tools can be designed to benefit from existing standardised tools and expand on them to fit specific requirements of CAs.

RQ 2. Is standardized usability questionnaire such as UMUX-Lite (Lewis, Utesch, & Maher, 2013) enough to inform about perceived usability of a chatbot?

The goal of this exploratory study is to identify a list of key factors that shape the perception of usability in interaction with chatbots and propose direction for the development of the new tool based on the key factors and existing tools. To reach this goal present study aims to:

A. Review factors that play a role in chatbots' perceived usability by examining present literature.
B. Examine the reliability and validity of UMUX-Lite (Lewis, Utesch, & Maher, 2013) in measuring perceived usability of a chatbot.
C. Inform the development of a potential new tool that will be more suited for the assessment of chatbots' perceived usability.

Overall research can be divided into three phases: the systematic literature review, the online survey, and the interaction test. Following sections will discuss these phases in more details. Section 2 will discuss the systematic literature review and its results. Section 3 will cover the methods and results (expert and end-user results) of the online survey. Section 4 describes the interaction test phases (pre-test, test, post-test), methods, and results (UMUX-Lite, Comparative Analysis of Test Pairs, post-test survey, debrief). Section 5 is intended to present the final list of the factors that were vetted based on the online survey and the interaction test. Finally, Section 6 will discuss results and general findings of the study.

2. SYSTEMATIC LITERATURE REVIEW

The literature review was conducted to explore aspects that influence a user's perception of the chatbot. It was a pivotal part of the study as discovered factors were planned to be used in the online survey and the interaction test. Subsequently, once the list was formed, it was presented to experts and general end-users by means of an online survey and to participants in an interaction test. As each group has a different perspective and motive towards chatbots, it can help to identify factors that are important in shaping perceived usability of the user.

2.1 Methods

The systematic literature review was qualitative and followed the phenomenological method (Randolph, 2009). Cooper's (1988) Taxonomy of Literature Reviews was used to plan and create a frame for the literature review. The rationale was to identify factors that contribute to the perceived usability in interaction with CA. Thus, the review focused on the studies that include findings and theories on factors that can potentially influence perceived usability of CA and studies that include assessment methods that might inform about criteria used during the assessment. Data was collected from journal databases (Scopus, Web of Science), Google Scholar and Google search engines, industry leaders' websites (Google, Amazon), and subject-specific professional websites. Search terms used in the search queries are: "conversational interface", "conversational agent", "chatbot", "interaction", "quality", "satisfaction". Considering the ever-evolving nature of the study subject and rapid growth in the interest towards CAs, database search included articles within the past ten years. Search query was also configured to limit results to journal articles and conference proceedings. A database search on Scopus was initially configured to exclude studies about virtual assistants and voice-controlled devices as they have qualities that do not coincide with another CAs and specifically with chatbots. However, further review of the literature proved that such sources also could contribute to the list of factors. Thus, the rest of the review did not exclude studies about virtual assistants and voice-controlled devices. More details of the database search are reflected in the search query syntaxes included in Appendix A.

Data from Scopus and Web of Science were compiled in reference management software (EndNote X8, Boston, MA, USA) and duplicates were removed. After duplicates were removed, abstracts of the documents were screened. Documents that described technical aspects and measurements of the CA and records that did not inform about CA's interactive characteristics were excluded from the list of eligible papers. The items in the literature review were analysed

based on the abstract and introduction to establish compliance with inclusion and exclusion criteria. After the screening, the remaining full text of the documents was assessed to determine the eligibility of the papers. When the list of qualified materials and factors were identified, they were examined more closely to create an initial pool of factors. During this stage, documents eligible for the study were used to form items for the list of factors. A PRISMA 2009 Flow Diagram was used to present the flow throughout the different stage of a systematic review (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009) (see Figure 1).

The review started from research in journal databases. However, it was not as fruitful as expected, so the research continued in the references of the relevant papers and Google Scholar. Only one study, written by Radziwill & Benton (2017), attempted to compile a comprehensive list of factors. The list from this study was used as a groundwork to the later developed list in Figure 2. Each factor from Radziwill and Benton's (2017) list was reviewed for relevance to the current study and tracked to its source. In the review process, new factors and sources were added, and some were excluded. Factors that were excluded were mainly the ones that affected the overall quality of the CA but were not that relevant to the perceived quality of the interaction.

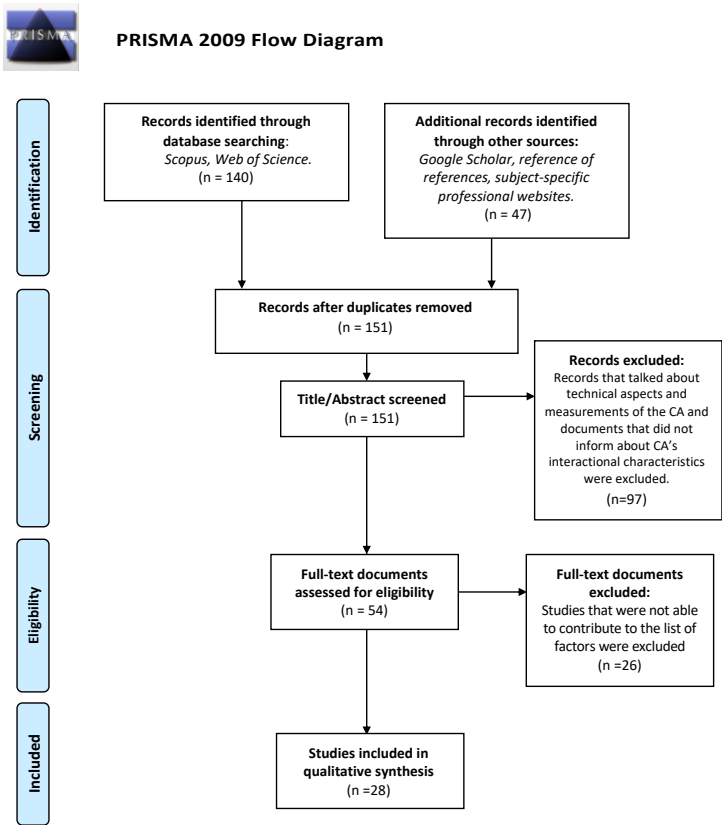


Figure 1. PRISMA 2009 Flow Diagram

2.2 Results

The rationale behind systematic literature was to identify and compile a list of factors that had contributed to the perceived usability in interaction with CAs. The literature review showed that many factors could affect the perception of the users. Forty-seven factors were identified from the literature. As a result, 47 factors that might affect the perception of interaction with CA were identified. Later similar factors were grouped in one broader factor, reducing number of factors to 27. These factors were compiled into a list along with their interpretations (Figure 2). Later, these factors were used to learn the opinion of developers, participants and end-users.

List of key factors that affect users' perception of usability

1. **Response Time.** Ability of the chatbot to respond timely to users' requests (Amazon, n.d.-b).
2. **Multi-thread conversation.** Ability of the chatbot to recognise and process multiple parallel topics simultaneously (Staven, 2017)
3. **Maxim of quantity.** Ability of the chatbot to respond in an informative way without adding too much information (Gnewuch et al., 2018; Google, 2017).
4. **Maxim of quality.** Ability of the chatbot to avoid false statements/information (Gnewuch et al., 2018; Google, 2017).
5. **Maxim of manners.** Ability of the chatbot to make it is purpose clear and communicate without ambiguity (Gnewuch et al., 2018; Google, 2017).
6. **Maxim of relation.** Ability of the chatbot to provide the relevant and appropriate contribution to people needs at each stage (Gnewuch et al., 2018; Google, 2017).
7. **Appropriate degrees of formality.** Ability of the chatbot to use appropriate language style for the context (Kirakowski et al., 2009).
15. **Perceived ease of use.** The degree to which a person believes that to interact with a chatbot would be free of effort (Van Eeuwen, 2017).
16. **Engage in on-the-fly problem solving.** Ability of the chatbot to solve problems instantly on the spot (Solomon, 2017).
17. **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 (Google, 2017; Kirakowski, Odonnell, & Yiu, 2009; Kuligowska, 2015).
18. **Breadth of knowledge.** Ability to exhibit knowledge that it is out of its immediate domain during a conversation (Cohen & Lane, 2016; Kirakowski, Odonnell, & Yiu, 2009; Vetter, 2002)
19. **Initiative.** Ability of the chatbot to initiate conversation (or offer cues) for further discussion by presenting its functionality, offering suggestions etc. (Amazon, n.d.-c; Google, 2017; Kirakowski et al., 2009; Kuligowska, 2015; Staven, 2017)
20. **Personality.** Ability of the chatbot to convey personality, warmth, and authenticity by providing greetings, self-introductory, empathy, information etc. (Amazon, n.d.-a; Kirakowski et al., 2009; Kuligowska, 2015; Lee & Choi, 2017; Solomon, 2017)
21. **Interaction enjoyment.** Chatbot is perceived as enjoyable and engaging to operate regardless of whether it provides in terms of information (Lee & Choi, 2017; Van Eeuwen, 2017).

List of key factors that affect users' perception of usability	
8. **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 (Google, 2017).	22. **Read and respond to moods of human participant.** Ability of the chatbot to appropriately recognise the mood of the user from its utterances and respond accordingly (M. O. Meira, 2015; Solomon, 2017).
9. **Visual Look.** The outward appearance of a chatbot's dialogue box, avatar, font etc. (Kuligowska, 2015).	23. **Sensitivity to safety and social concerns.** Ability of the chatbot to recognise, respond to safety or social concern and refer a user to helpline if needed (Miner et al., 2016).
10. **Voice Tone.** Spoken expressiveness (inflexion, emotional information through tone) and the accuracy of the text-to-speech function of the chatbot (Kuligowska, 2015; Pauletto et al., 2013).	24. **Meets diversity needs.** Ability of the chatbot to meet needs of users independently form their health conditions, well-being, age etc. (Radziwill & Benton, 2017).
11. **Integration with the website.** Position in the website and visibility of the chatbot (all pages/specific pages, floating window/pull-out tab/embedded etc.) (Kuligowska, 2015).	25. **Trustworthiness.** Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage (Hertzum, Andersen, Andersen, & Hansen, 2002; Lee & Choi, 2017).
12. **Graceful responses in unexpected situations.** Ability of the chatbot to gracefully handle unexpected input, communication mismatch and broken line of conversation (Amazon, n.d.-a; Cohen & Lane, 2016; Ramos, 2017; Van Eeuwen, 2017; Wilson, Daugherty, & Morini-Bianzino, 2017)	26. **Process tracking and follow up.** Ability of the chatbot to inform and update users about the status of their task in progress (Van Eeuwen, 2017).
13. **Recognition and facilitation of users' goal and intent.** Ability of the chatbot to recognize user's intent and guide the user to its goal (Coniam, 2014; Ramos, 2017; Van Eeuwen, 2017a; Wilson et al., 2017).	27. **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 (Applin & Fischer, 2015; Van Eeuwen, 2017).

Figure 2. List of key factors that affect users' perception of usability.

3. ONLINE SURVEY

An online survey was designed to compare the opinions of the experts and end-users. Opinions of the experts and users were used to formalise the list of key factors essential for the evaluation of perceived usability.

3.1 Methods

3.1.1 Participants

The survey respondents were 20 European participants and comprised of 11 end-users and 9 experts. Two of the 20 participants were female the rest were male. The average age of respondents was 38 with a standard deviation of 9.4 years An Italian company supported this project in kind (UserBot.ai) and they helped in the diffusion of the survey. Thus, about 95% of

the participants were Italian. Five out of nine experts have two or fewer years of experience with chatbots, three have two to five years, and one has more than five years of experience. All end-users hold at least a high school education and use chatbots at least once a week. Each participant has been presented a consent form that included information about the study and the contact details of researchers at the beginning of the survey. Respondents who answered to less than ten factors were excluded from the sample, leaving 9 expert and 8 end-user data useful for the analysis.

3.1.2 Apparatus

The primary tool used in this part of the study was an online survey. The online survey was created through an online survey platform (Qualtrics, Provo, UT, USA) that can generate anonymous links which users can use to access the survey and fill it in. The survey consisted of: (a) consent form with study information, (b) personal and professional information, (c) 27 factors with descriptions on 7-point Likert scale with statements ranging from "Strongly disagree" to "Strongly Agree", (d) sorting task in which 27 factors can be categorised into 'core', 'dependent' and 'marginal' aspects (experts only), (e) survey feedback, (c) contact information in case respondent is willing to participate in further research (Appendix B). The reinforced an informed consent form that was inserted to the beginning of the of the survey (Appendix B). Consent form template was retrieved from the University of Twente website and modified to meet the needs of this study.

3.1.3 Design and Procedures

The survey was exploratory and was administered online. The aim of the survey was mainly to get the opinion of the experts along with the interaction test participants. However, the survey was also open for the end-users who might want to fill in the survey. The survey included explanations for the factors and item descriptions to make sure respondents understood what was expected of them. The survey was configured in a way to show relevant questions based on whether the user is expert or end-user. The main difference in survey composition for these two groups was that end-users were not asked professional questions and were not presented with the sorting task. For more details on the individual survey items and structure refer to Appendix B.

3.1.4 Data Analysis

Data entered to the online survey was gathered and exported through survey management software (Qualtrics, Provo, UT, USA). Next data was cleaned and prepared for the analysis. Analysis started from 27 statements on a 7-point Likert scale that represented each factor. The

data file was split into two, based on the status of the respondents (expert or end-user). Consequently, the statistical output was generated separately for each group. Median scores for each factor were calculated to discriminate the statements for which response results fall above and below 50 per cent of the scale. Interquartile ranges (IQRs) were used to interpret the spread of the data and to estimate the level of agreement per factor (Polisena et al., 2019).

Analysis of sorting task started with Fleiss' Kappa method which is used to measure consensus between raters (experts) and rules out consensus by chance. In sorting task participants could assign factors into three categories; "marginal aspects", "core aspects", or "dependent aspects". In Fleiss' Kappa method it is important that each case is evaluated the equal amount of times. Thus, Fleiss' Kappa analysis was conducted with 7 experts since one expert did not fill in the sorting task at all and another one missed first 10 questions. Fleiss' Kappa was calculated using an online calculator (Randolph, 2008). Further, to extrapolate factors with the strongest consensus, factors and categories were placed in rows and columns, forming a matrix. The number of times raters placed a factor to the specific category was inserted into the corresponding cell. Then each factor was labelled based on which category had more counts. If there was a factor with the same number of counts for two categories, it was compared to the results from scale to decide which category it belongs to.

3.2 Results

3.2.1 Experts

Along with the demographic questions, scales and sorting questions, experts were also asked about their opinion on the following questions:

- On the basis of your knowledge, in your field which are the most frequently used approaches applied by *other companies* to test quality perception of the end-users in interaction with chatbot?
- Which approaches do *you* usually use to test the quality perception of the end-users in interaction with chatbot?
- Please briefly list, what are the main aspects which you take into account when *you* design chatbots?

Majority of experts reported (a) observation of user's interaction, (b) remote log analysis, (c) usability test with small groups to be most commonly used approaches applied by other companies to test the end-users perceived quality of the interaction with a chatbot. Whereas, majority experts noted personally using (a) observation of the user's interaction, (b) conversation

analysis, (c) usability test with small groups, (d) interviews as an assessment method in their company. Before experts moved to the factors scale, they answered the question: what are the main aspects you take into account when you design chatbots? Experts replied the following:

- *Expert 1:* User experience, clear conversation flows, AI training.
- *Expert 2:* Efficiency, accuracy in the answers, the ability to understand the needs of the user, speed in providing answers.
- *Expert 3:* Usability according to end user's behaviour.
- *Expert 4:* Available data for training, the domain of application, available time for realization, available computational power.
- *Expert 5:* Mobility Interaction Cleaned Design.
- *Expert 6:* Topic/general purpose, user category, UX design.
- *Expert 7:* The personality, the ability to give an answer to "non-conventional" words or provocations, the possibility to establish an emphatic conversation, the ability to avoid loop answers.
- *Expert 8:* We take care to guide the user with buttons, hints and try to follow users' intents in natural language processing with conversational hints.

The answers of the experts reflect differences in their perspectives and aspects they find important. These differences seem to originate from the roles they hold (i.e. developer, head of operations, researcher, designer). To determine which 27 factors expert reached strong consensus—criteria for strong agreement and disagreement was established for the responses on a 7-point Likert scale. Factors with a median of ≤ 3 (i.e., factors that scored from "Somewhat disagree" to "Strongly disagree") and IQR in the range from 3 to 1 were regarded to have reached concurrence on a strong disagreement. Factors with a median ≥ 5, (i.e., factors that scored from "Somewhat agree" to "Strongly agree") with an IQR within range of 5 and 7 were considered to have reached consensus on a firm agreement with the suggested factor. The factors which IQRs did not fall between the range of 5 and 7 but had a median of ≥ 5 were excluded from the list in Figure 3. Median and IQR score for each factor is reported in Table 1.

Online survey results for each factor statement					
		Experts		***End-users***	
No.	***Factor***	***Median (IQR)***	***No. of participants (%)***	***Median (IQR)***	***No. of participants (%)***
F1	*Response time*	6 (5-6.5)	9 (100%)	6 (5-6.75)	8 (100%)
F2	*Multi-thread conversation*	5 (4-6)	9 (100%)	5 (4-6)	8 (100%)
F3	*Maxim of quantity*	6 (5-6)	9 (100%)	6 (4.25-6.75)	8 (100%)
F4	*Maxim of quality*	7 (7-7)	9 (100%)	6 (4.5-7)	8 (100%)
F5	*Maxim of manners*	7 (6-7)	9 (100%)	6 (4.25-6)	8 (100%)
F6	*Maxim of relation*	6 (6-7)	9 (100%)	6 (5-6)	8 (100%)
F7	*Appropriate degrees of formality*	6 (5.5-6)	9 (100%)	5 (3.5-5.75)	8 (100%)
F8	*Reference to what is on the screen*	7 (6-7)	9 (100%)	6 (4.25-6.75)	8 (100%)
F9	*Visual Look*	6 (4-6)	9 (100%)	4.5 (3-5.75)	8 (100%)
F10	*Voice Tone*	5 (4-6)	9 (100%)	4.5 (4-6)	8 (100%)
F11	*Integration with the website*	6 (6-7)	8 (87.5%)	5 (5-6)	7 (87.5%)
F12	*Graceful responses in unexpected situations*	6 (6-7)	8 (87.5%)	6 (6-7)	7 (87.5%)
F13	*Recognition and facilitation of users' goal and intent*	7 (6-7)	8 (87.5%)	6 (5-7)	7 (87.5%)
F14	*Variation of responses*	5 (4.25-5.75)	8 (87.5%)	4 (3-5)	7 (87.5%)
F15	*Perceived Ease of Use*	6 (4-6.75)	8 (87.5%)	6 (5-6)	7 (87.5%)
F16	*Engage in on-the-fly problem solving*	6 (5.25-7)	8 (87.5%)	6 (5-6)	7 (87.5%)
F17	*Themed discussion*	6 (5.25-6.75)	8 (87.5%)	5 (4-7)	7 (87.5%)
F18	*Breadth of knowledge*	5 (4-5)	8 (87.5%)	5 (4-7)	7 (87.5%)
F19	*Initiative*	5 (4.25-6)	8 (87.5%)	5 (3-5)	7 (87.5%)
F20	*Personality*	6 (5-7)	8 (87.5%)	5 (4-6)	7 (87.5%)
F21	*Interaction enjoyment*	4.5 (2.50-5.75)	8 (87.5%)	4.50 (2.50-5.25)	6 (75%)
F22	*Read and respond to moods of human participant*	5.5 (5-6)	8 (87.5%)	5 (4-6)	6 (75%)
F23	*Users' privacy and ethical decision making*	7 (6-7)	8 (87.5%)	7 (6-7)	6 (75%)
F24	*Sensitivity to safety and social concerns*	6.5 (5-7)	8 (87.5%)	5.5 (3.75-6.25)	6 (75%)
F25	*Meets neurodiverse needs*	6 (5-6.75)	8 (87.5%)	5 (3.75-6.25)	6 (75%)
F26	*Trustworthiness*	6 (5-6)	8 (87.5%)	4.50 (3.75-6.25)	6 (75%)
F27	*Process facilitation and*	6.5 (6-7)	8 (87.5%)	6.50 (5-7)	6 (75%)

Online survey results for each factor statement
follow up

Table 1. Online survey results for each factor statement.

Analysis of median scores and IQRs of 27 factors on 7-point Likert scale showed agreement on 19 factors. Analysis of the sorting task revealed that not all the factors that experts strongly agreed on in response to the scales are indeed important. On the contrary, some factors that experts did not strongly agree in their responses to the scales are perceived to be important. Inter-rater reliability for the sorting task showed slight agreement between raters, k= 0.12 (7 items) with 42% overall agreement. Which indicates that results obtained from the sorting task are reliable enough to be used in the analysis. Subsequently, results of the sorting task helped to refine the list of 16 factors that experts find important in the assessment of chatbots perceived usability (Figure 3).

List of factors on which experts reached consensus			
1.	Response Time	2.	Maxim of quality
3.	Maxim of manners	4.	Maxim of relation
5.	Appropriate degrees of formality	6.	Reference to what is on the screen
7.	Integration with the website	8.	Recognition and facilitation of users' goal and intent
9.	Variation of responses	10.	Perceived ease of use
11.	Engage in on-the-fly problem solving	12.	Themed discussion
13.	Personality	14.	Users' privacy and ethical decision making
15.	Trustworthiness	16.	Process facilitation and follow up

Figure 3. List of factors on which experts reached consensus

3.2.2 End-users

End-users were also presented 27 factors with descriptions on the 7-point Likert scale. Selection procedure of the factors that will be included in the list was similar to the experts. Factors with a median ≥ 5, (i.e., factors that scored from "Somewhat agree" to "Strongly agree") with an IQR within range of 5 and 7 were considered to have reached consensus on a firm agreement and were added to the list. The factors which IQRs did not fall between the range of 5 and 7 but had a median of ≥ 5 were excluded from the lists in Figure 4. Median and IQR score for each factor is reported in Table 2. Analysis of median scores and IQRs from Table 1 showed that end-users reached firm consensus on 11 factors that are reported in Figure 4.

List of factors on which end-users reached consensus			
1.	Response time	2.	Recognition and facilitation of users' goal and intent
3.	Maxim of relation	4.	Perceived Ease of Use
5.	Appropriate degrees of formality	6.	Engage in on-the-fly problem solving
7.	Reference to what is on the screen	8.	Users' privacy and ethical decision making

9. Integration with the website
10. Process facilitation and follow up
11. Graceful responses in unexpected situations

Figure 4. List of factors on which end-users reached consensus.

4. INTERACTION TEST

The interaction test was designed and conducted for the following two reasons: (a) get perspective of the users with recent experience (reduce memory biases) and (b) check the reliability of the UMUX-Lite (Lewis, Utesch, & Maher, 2013) as a measurement of satisfaction during the interaction with chatbots.

4.1 Methods

4.1.1 Participants

The interaction test involved 16 participants in total. Two out of 16 participants were pilot participants. Records of the first pilot participants were lost due to a technical issue, therefore, they were not included in the study leaving a total of 15 participants of which seven were female and eight male. Participants were students of the University of Twente and were recruited through the test-subject pool system or via convenience sampling. There was no screening of the participants each participant that signed up was admitted for participation, however, signees were notified beforehand that they would require a good command of English before they signed up. Subjects recruited through test-subject pool system received credit points for their participation.

The mean age of participants was 23 with a standard deviation of 3.8 years. Eight of the participants were male, and seven female and they were from the Netherlands, Germany, Britain, Colombia, India, USA, Mexico, and Spain. All participants have obtained at least a high school diploma and have at least intermediate level of English. Majority of participants were from Engineering and Psychology background however there was one person with Philosophy and one with Computer Science background. Even though all participants stated that they were familiar with chatbots to a certain degree, five participants also stated that they did not use chatbots before.

4.1.2 Apparatus

The experiment was conducted in a quiet lab room. Lab room included essential equipment as desk, chairs, desktop computer and a webcam. To administer the surveys this part of the study also used online research platform (Qualtrics, Provo, UT, USA). Usability test software was used to record the interaction test as a whole. Only data from debriefing was used

in the qualitative analysis of the data. Materials used during the interaction test were the informed consent form, test script, pre-test survey, task cards, UMUX-Lite (Lewis et al., 2013) and post-test survey. Materials can be found in Appendix C.

4.1.3 Design and Procedures

The interaction test had two conditions "Chatbot" and "Web Navigation." The experiment had a 2x6 within-group design. Tasks were paired, which means that a participant executed the same scenario once with chatbot and once with web navigation. Moreover, tasks were randomised and efforts were taken to ensure that paired tasks did not follow each other. During the interaction test, participants were instructed to think aloud and to fill in the UMUX-Lite scale (Lewis et al., 2013) after the end of each task. Scenarios were divided into three categories: embodied chatbots (2 tasks), text chatbots (2 tasks), messenger chatbots (2 tasks).

The interaction test can be divided into three phases: pre-test, test, and post-test. Pre-test phase started by greeting the participant, giving study information and shortly explaining chatbots if the participant is unsure what is a chatbot. After all the essential information has been verbally presented, the participant was given two copies of the consent form. One copy had to be signed and returned and the participant kept the second copy. After the participant signed the consent form, he or she was presented with a pre-test survey on a computer screen.

Pre-test phase was followed by the test phase where participants had to perform a series of tasks with chatbots and web navigation. First, a more detailed explanation of the test was given. Participants were briefed on (a) how to use tools, (b) how to fill in the survey, (c) how to act, (d) how to think aloud, and (e) how to interact with test administrator. To clarify what is expected of the participant, the think-aloud protocol was demonstrated to the participant as an example. The administrator also emphasised that there is no right, or wrong answer and every comment of the participant is a valuable contribution.

When a participant had no questions and was ready to start, the test administrator started recording the test. Before actual test scenarios were given to the participants, they were asked to interact with a chatbot called Mitsuku (https://www.pandorabots.com/mitsuku/) to practice for 1-2 minutes. Mitsuku is a chatbot that was designed to imitate a teenager and converses with users on general topics. Once the participants were ready to move onto the main tasks, the test administrator presented the first task card to the user. Participants were encouraged to ask questions if the scenarios were unclear. Task cards were organised in two decks based on conditions and were shuffled before each participant or when two same tasks followed each

other. Right after participants notified of the completion of the task, they filled in UMUX-Lite (Lewis et al., 2013) and marked the task code that was given on the task card. Participants repeated the procedure until all the tasks were finished.

Post-test phase consisted of a debrief and post-test questionnaire (Appendix C). After the last task was completed, the participants were asked to answer questions about the experience they had today interacting with chatbots (Appendix C). All participants answered the same set of questions. Test administrator stopped recording after debriefing. Debrief was followed by the post-test questionnaire where participants had to express their agreement about statements that represented each factor (Appendix C). Participants were encouraged to ask questions if any statement was unclear. When participants were done with filling in the survey, they were thanked for participation and were notified about the end of the experiment.

4.1.4 Data Analysis

Data entered into the pre-test, post-test survey and UMUX-Lite was gathered and exported through online survey platform (Qualtrics, Provo, UT, USA). Afterwards, data was cleaned and prepared for analysis. The main purpose of the data collected from the pre-test survey was to establish sample demographic.

UMUX-Lite scores were calculated with the formula provided by Lewis et al. (2013) that provide scores corresponding with SUS scores. Mean UMUX-Lite scores per chatbot and website were computed to explore their level of usability in relation to each other. Since, applied formula produced scores corresponding to SUS, obtained scores can be regarded as SUS scores.

Sourcing data from 446 studies and more than 5,000 individual SUS responses, Sauro and Lewis (2012) found the overall mean score of the SUS to be 68 with a standard deviation of 12.5. Further, they proposed curve grading scale (CGS) for SUS scores. CGS is in the range from F (absolutely unsatisfactory) to A+ (absolutely satisfactory). A grade above and including C is considered to be acceptable. Grades and corresponding SUS scores are listed below:

- Grade F (0–51.7)
- Grade D (51.8–62.6)
- Grade C– (62.7–64.9)
- Grade C (65.0–71.0)
- Grade C+ (71.1–72.5)
- Grade B– (72.6–74.0)
- Grade B (74.1–77.1)

- Grade B+ (77.2–78.8)
- Grade A– (78.9–80.7)
- Grade A (80.8–84.0)
- Grade A+ (84.1–100)

Additionally, Sauro and Lewis (2012) also found that public facing large scale websites showed an average score of 67, interactive voice response (IVR) system showed an average score of 79.9 and a combination of web-based IVR showed score average score of 59.2 (Sauro, 2011). CGS will be used to review mean SUS scores for websites and chatbots.

Moreover, a paired-samples t-test was conducted to see if the observed difference in mean SUS scores for chatbot and website conditions are significant. Cronbach's alpha was used to assess the inter-item reliability of the UMUX-Lite. To calculate Cronbach's alpha all UMUX-Lite scores in chatbot condition for Item 1 and Item 2 were compiled in the new dataset. Then reliability analysis was performed to get Cronbach's alpha.

Like the online survey, the 27 statements about key factors for the usability of chatbots was presented in the post-test phase of the interaction test, and IQRs were calculated to establish an agreement (Polisena et al., 2019). Participants' answers to debrief questions (Appendix C) were noted down from the test recordings and qualitatively analysed. For each question, answers that overlapped were grouped and reported as one general theme.

4.2 Results

4.2.1 UMUX-Lite and Comparative Analysis of Test Pairs

Reliability of α=0.885 (15 items) was found for UMUX-Lite in chatbot condition. For website condition reliability coefficient was α= 0.902 (15 items). For both conditions, findings are in line with Lewis et al. (2013). Results of paired t-test (Table 2) showed significant differences in the mean SUS scores for Inbenta ($M=28.16$, $SD=22.63$, $t(14)=4.82$, $p=0.001$) and Australian Tax Office ($M=-16.61$, $SD= 17.54$, $t(14)=-3.66$, $p=0.003$) pairs. Although, a t-test showed that there was a difference for the chatbots Hipmunk, Finn, Yoko, Veronica, Julie, the difference was not significant (Table 2).

Hipmunk pair has grades close to each other, with chatbot(C) showing score barely above average and website below the average. Finnair chatbot(B–) and website(A–) both have acceptable SUS scores. Yoko has a mean score below average(C-), meanwhile, its website alternative has usability grade above average(B). Inbenta's Veronica has the lowest usability score (50.70) among all the test articles which can be interpreted as completely unsatisfactory(F).

However, Inbenta's website was rated to be satisfactory (B+). Amtrack's chatbot(B) and website(B+) had almost the same level of satisfaction, with chatbot being rated a bit higher. Australian Tax Office chatbot has the highest grade(A) among all test articles and is far above average, whereas, the website is barely passing average mark(C). Overall, chatbots have higher standard deviation scores compared to websites, which indicates that there is more variance in the opinions of the raters in a chatbot condition than in website. Nevertheless, websites of Hipmunk, Google and Australian Tax Office also show variance in responses.

Mean UMUX-Lite scores, standard deviations per task pairs and paired t-test results

Tasks	***Website***		***Chatbot***		***t-test***	
	Mean (CGS)	***SD***	***Mean (CGS)***	***SD***	***Mean difference***	***t(df) /p-value***
Hipmunk *(Hipmunk)*	60.09 (A–)	21.36	68.40 (C)	22.12	*–8.30*	*t(14)=–1.07 p=0.303*
Finn *(Finnair)*	79.95 (A–)	5.36	73.84 (B–)	20.51	*6.11*	*t(14)=1.33 p=0.205*
Yoko *(Toshiba)*	74.90 (B)	18.17	64.42 (C–)	15.41	*10.47*	*t(14)=1.55 p=0.142*
Veronica *(Inbenta)*	78.87 (B+)	11.87	50.70 (F)	21.75	*28.26*	*t(14)=4.82 p=0.001*
Julie *(Amtrack)*	73.45 (B–)	9.30	74.17 (B)	11.73	*–0.72*	*t(14)=–0.2 p=0.845*
Alex *(Australian Tax Office)*	66.23 (C)	18.19	82.84 (A)	5.95	*–16.61*	*t(14)=–3.66 p=0.003*

Table 2. Mean UMUX-Lite scores, standard deviations per task pairs and paired t-test results.

4.2.2 Post-test survey

Median and IQR were computed for responses per each factor. Median and IQR score for each factor is reported in Table 3. Next, inclusion criteria were applied for each factor. Factors with a median ≥ 5, (i.e., factors that scored from "Somewhat agree" to "Strongly agree") with an IQR range of 5 and 7 were considered to have reached consensus on a firm agreement and were added to the list. Analysis of participants' answers to 27 usability factors on 7-points Likert scale showed a strong consensus on 16 factors listed in Figure 5.

List of factors on which interaction test participants reached consensus.			
1.	Response time	2.	Recognition and facilitation of users' goal and intent
3.	Multi-thread conversation	4.	Variation of responses
5.	Maxim of quantity	6.	Ease of Use
7.	Maxim of quality	8.	Engage in on-the-fly problem solving
9.	Maxim of manners	10.	Users' privacy and ethical decision making
11.	Maxim of relation	12.	Meets neurodiversity needs
13.	Reference to what is on the screen	14.	Trustworthiness
15.	Integration with the website	16.	Process facilitation and follow up

Figure 5. List of factors on which interaction test participants reached consensus.

Post-test survey results for each factor			
No	***Factors***	***Median (IQR)***	***No. of participants (%)***
F1	*Response time*	6 (6-7)	15 (100%)
F2	*Multi-thread conversation*	6 (5-7)	15 (100%)
F3	*Maxim of quantity*	7 (6-7)	15 (100%)
F4	*Maxim of quality*	7 (6-7)	15 (100%)
F5	*Maxim of manners*	6 (5-7)	15 (100%)
F6	*Maxim of relation*	6 (6-7)	15 (100%)
F7	*Appropriate degrees of formality*	6 (4-6)	15 (100%)
F8	*Reference to what is on the screen*	6 (5-6)	15 (100%)
F9	*Visual Look*	4 (3-5)	15 (100%)
F10	*Voice Tone*	6 (4-6)	15 (100%)
F11	*Integration with the website*	7 (6-7)	15 (100%)
F12	*Graceful responses in unexpected situations*	6 (3-6)	15 (100%)
F13	*Recognition and facilitation of users' goal and intent*	7 (6-7)	15 (100%)
F14	*Variation of responses*	6 (5-7)	15 (100%)
F15	*Perceived Ease of Use*	7 (6-7)	15 (100%)
F16	*Engage in on-the-fly problem solving*	7 (5-7)	15 (100%)
F17	*Themed discussion*	6 (4-7)	15 (100%)
F18	*Breadth of knowledge*	4 (3-4)	15 (100%)
F19	*Initiative*	5 (3-5)	15 (100%)
F20	*Personality*	6 (4-6)	15 (100%)
F21	*Interaction enjoyment*	5 (4-6)	15 (100%)
F22	*Read and respond to moods of human participant*	3 (2-5)	15 (100%)
F23	*Users' privacy and ethical decision making*	6 (5-6)	15 (100%)
F24	*Sensitivity to safety and social concerns*	5 (4-5)	15 (100%)
F25	*Meets neurodiverse needs*	6 (5-7)	15 (100%)

Post-test survey results for each factor			
F26	*Trustworthiness*	6 (5-7)	15 (100%)
F27	*Process facilitation and follow up*	7 (7-7)	15 (100%)

Table 3. Post-test survey results for each factor.

4.2.3 Debrief results

Participants' answers to the questions given during debriefing are in line with the factors on which they reached agreement consensus on in Figure 4. Analysis of the answers revealed some patterns and repeated themes. Discovered patterns for each question are reported in the next paragraphs.

Q1: What are the things that you did not like during your interaction with chatbot?

Seven out of 15 participants stated that they do not like to rephrase a question several times in order to get results. Some participants also mentioned that they do not like adapting their conversational style to the chatbot, by simplifying their input. They prefer when chatbot understands their full sentences because it feels more natural. Three participants mentioned that they did not like that it took them longer to accomplish their tasks with chatbots than they expected, and one commented that "it is supposed to be a faster way". Two participants mentioned that they did not like chatbots starting the conversation without them initiating it as they either felt obligated to answer or felt overwhelmed with the provided information. Three test participants noted their inability to find chatbot at all or spending too much time to search for it; they would prefer chatbot to be more visible like some chatbots they have tested. It is important to note that two participants that could not find chatbot at all were the ones that were not well familiar with chatbots. Lastly, participants complained about chatbot outputs having too much information and providing to broad or generic answers.

Q2: What are the things that you liked during your interaction with chatbot?

There are a couple of central themes in the answers for this question. Respondents pointed out that they appreciated fast instantaneous responses and directness of interaction in comparison to graphical user interfaces (GUIs). Concerning directness, they often mentioned that they just needed to ask, and they got information. A couple of people also mentioned that they liked it when chatbot exhibited personality or other anthropomorphic features.

Q3: Which chatbot did you enjoy interacting with the most? Why?

The chatbot that participants enjoyed the most was chatbot Alex on Australia's Tax Office website (Appendix C) seven participants suggested this chatbot. They mentioned that it was effortless and gave them the information they needed. Next with five responses was

Hipmunk's chatbot (Appendix C), participants liked it because, it was entertaining, gave various suggestions and feedback, predicted users' questions, remembered previous input, and had visuals. Finnair's chatbot (Appendix C) came third with four mentions, and participants liked it for the same reasons as Alex. Amtrack's Julie (Appendix C) and Inbenta's Veronica (Appendix C) also got one mention each. None of the users mentioned Toshiba's chatbot Yoko (Appendix C).

Q4: With which chatbot did you not enjoyed interacting? Why?

Participant did not like interaction with Inbenta's Veronica, because users could not get answers even after rephrasing their input, they felt like she did not understand them. Toshiba's Yoko and Amtrack's Julie both had three mentions, they did not like Yoko because she provided many links instead of a direct answer and had a synthetic voice and they did not like Julie because she was hard to find and redirected users to the pages instead of giving a direct answer. Participants also mentioned Finnair's and Hipmunk's chatbots. None of the users mentioned Alex from the Australian tax office.

Q5: Can you tell me if you prefer more to use chatbots or website navigation? In which situations you would prefer using chatbots to the standard navigation?

Most of the participants stated that they prefer standard navigation. Since they are already familiar with the basic structure of the website and usually can easily find the information they want. However, many participants mentioned that they would be more likely to use chatbots after their experience during the interaction test. Correspondingly they stated that it depends on the chatbot and the situation. Participants said that they would use chatbot if they are on an unfamiliar website or the large governmental website with much information. Additionally, they said they would use chatbot if they do not find needed information through the website. Some participants told that they would use chatbot if they have a specific question in mind. On the contrary, some participants remarked that they would prefer to use chatbot when they do not have an idea of what they need. Furthermore, participants mentioned that they would use chatbot if they need simple information, if the information is too complicated like flight search or repair instructions, they would prefer a website.

Q6: In your opinion, what qualities or features were missing, what features could be improved or added?

Participants told that they would like chatbot to offer more cues, for example, they would like chatbot to give cues and tell how to phrase questions like Finnair chatbot. Meanwhile, some

preferred chatbot to be more interactive and less robotic with anthropomorphic features (e.g. personality, avatar, voice, use of emojis) some users told that they would prefer chatbots to have less anthropomorphic features. Group of participants also said that they would prefer chatbots to be more visible (e.g. pop-up when opening the page) and have more visuals.

5. FINAL LIST OF FACTORS

To determine the final list of factors, data from the online survey and post-test survey was compiled together. This data included results for 27 usability factors on 7-points Likert scale filled in by experts, end-users and the interaction test participants. Similarly, to the procedure that was applied for online survey and the interaction test median and IQR for each factor was calculated and reported in Table 4.

Results from combined data for each factor			
No	***Factors***	***Median (IQR)***	***No. of participants (%)***
F1	*Response time*	6 (5-7)	32 (100%)
F2	*Multi-thread conversation*	5 (4-6)	32 (100%)
F3	*Maxim of quantity*	6 (5-7)	32 (100%)
F4	*Maxim of quality*	7 (6-7)	32 (100%)
F5	*Maxim of manners*	6 (5-7)	32 (100%)
F6	*Maxim of relation*	6 (6-7)	32 (100%)
F7	*Appropriate degrees of formality*	6 (5-6)	32 (100%)
F8	*Reference to what is on the screen*	6 (5.25-7)	32 (100%)
F9	*Visual Look*	5 (3-6)	32 (100%)
F10	*Voice Tone*	5 (4-5)	32 (100%)
F11	*Integration with the website*	6 (5-7)	30 (93.75%)
F12	*Graceful responses in unexpected situations*	6 (5.5-7)	30 (93.75%)
F13	*Recognition and facilitation of users' goal and intent*	7 (6-7)	30 (93.75%)
F14	*Variation of responses*	5 (4-6)	30 (93.75%)
F15	*Perceived Ease of Use*	6 (6-7)	30 (93.75%)
F16	*Engage in on-the-fly problem solving*	6.5 (5.75-7)	30 (93.75%)
F17	*Themed discussion*	6 (5-7)	30 (93.75%)
F18	*Breadth of knowledge*	4 (3-5)	30 (93.75%)
F19	*Initiative*	5 (3.75-5.25)	30 (93.75%)
F20	*Personality*	5.50 (4-6)	30 (93.75%)
F21	*Interaction enjoyment*	5 (4-6)	29 (90.62%)
F22	*Read and respond to moods of human participant*	5 (3-5.5)	29 (90.62%)
F23	*Users' privacy and ethical decision making*	6 (6-7)	29 (90.62%)
F24	*Sensitivity to safety and social concerns*	5 (4-6)	29 (90.62%)

Results from combined data for each factor			
F25	*Meets neurodiverse needs*	6 (5-6.5)	29 (90.62%)
F26	*Trustworthiness*	6 (5-6)	29 (90.62%)

Table 4. Results from combined data for each factor

After calculations selection criteria were applied to the calculated results. Factors with a median ≥ 5, (i.e., factors that scored from "Somewhat agree" to "Strongly agree") with an IQR within range of 5 and 7 were considered to have reached consensus on a firm agreement and were added to the final list of factors that were considered to be important by all study participants. Results show a strong consensus on 17 factors reported in Figure 5. From remaining ten factors nine reached agreement but had wider IQRs (between 3 and 6). For the factor "Breadth of knowledge," neither agreement nor disagreement was reached among the raters. Additionally, inspection of results across all three groups (experts, end-users and interaction-test participants) (see Figures 3, 4 and 5) showed complete consensus on 7 factors out of 17, which are marked with "*" and presented in Figure 5.

Review of the interaction test debrief showed that participant found linguistics capabilities of chatbot important. Majority of the participants noted that they do not like rephrasing their sentences multiple times or would like chatbot to understand their input even though there is a mistake. Even though one of the factors in the initial list "graceful responses in the unexpected situation", addresses this issue it only addresses the consequences of poor communication. Debrief showed that users find it important for chatbot to have good linguistic processing capabilities. Discussion linguistic processing abilities of the chatbot can be found in studies of Coniam (2014), Kuligowska (2015), and Kluwer (2011). Considering above mentioned information "linguistic flexibility of input" was added as the 18th factor and marked by "**". (see Figure 6).

The final list of factors with a strong consensus across all groups			
1.	Response time*	2.	Graceful responses in unexpected situations
3.	Maxim of quantity	4.	Recognition and facilitation of users' goal and intent*
5.	Maxim of quality	6.	Perceived Ease of Use
7.	Maxim of manners	8.	Engage in on-the-fly problem solving*
9.	Maxim of relation*	10.	Themed discussion
11.	Appropriate degrees of formality	12.	Users' privacy and ethical decision making*
13.	Reference to what is on the screen*	14.	Meets neurodiversity needs
15.	Integration with the website*	16.	Trustworthiness
17.	Process facilitation and follow up*	18.	Flexibility of linguistic input

Figure 6. Final list of factors with a strong consensus across all groups

6. DISCUSSION

The aim of this study was to (a) explore the factors that are essential for user satisfaction, (b) investigate UMUX-Lite scale as a usability measure of a chatbot, and to (c) propose a design approach for a potential usability questionnaire. This section will discuss results derived from this exploratory study, its limitations, and future recommendations.

6.1 Key Factors

As it was mentioned in Section 5, results showed a firm consensus for 17 out of 27 factors (Figure 6). The additional 18th factor was added to the final list based on the results of the debriefing. However, comparison of the consensus across the groups (Figure 3, 4, 5) showed complete consensus for 7 out of 17 factors. This could indicate that these 7 factors are more relevant for assessing perceived usability. However, given the limitations of the study specifically modest sample size and exploratory nature of this study, it was decided to take a more conservative approach and retain a list of 17 factors. Although this study was able to identify key factors that affect perceived usability in interaction with chatbots, these factors need further refinement. Future studies need to be conducted to ensure that the result of the presents study is valid.

6.2 UMUX-Lite as Measure of Usability and Its Limitations

The study results showed that UMUX-Lite in chatbot condition had strong inter-item reliability (α= 0.885). It is consistent with the internal reliability score for UMUX-Lite Lewis et al. (2013) found in his study (α=0.86). This indicates that items are correlated with each other, and they measure the same underlying construct–usability. Comparison of the means via paired t-test only found a significant difference for Inbenta and Australian Tax Office pairs. Which indicates that, to a certain extent, UMUX-Lite is sensitive to detect changes in the tested systems and can discriminate among chatbots with low and high usability. However, differences were not significant for rest of the pairs which question the extent to which UMUX-Lite valid and sensitive in measuring chatbots. Considering that several studies (Borsci, Federici, Bacci, Gnaldi, & Bartolucci, 2015; Lewis et al., 2013) found significant correlations between, SUS, UMUX and UMUX-Lite we can assume that these scales will produce similar results if they are used to measure satisfaction in chatbots.

There are some assumptions to why UMUX-Lite did not produce a significant difference for some pairs. One of the reasons for the insignificant difference might be a similar performance

of both chatbot and website. For example, comparative analysis (Section 4.2.1) showed that for the pairs Hipmunk, Finnair and Amtrak perceived usability for chatbot and website is on a similar level. Which might explain why the difference for this pairs were insignificant. It might have been possible to get more reliable results if pairs included one system that is known for good performance and one for bad, as in the study of Finstad (2010). However, the assumption that similar performance is the reason for insignificant difference does not hold itself completely. In the case of Toshiba pair, chatbot has a SUS score below average (64.42, C–) whereas website has a score above average (74.90, B) nonetheless results are still is insignificant.

Another assumption for insignificant differences might be the sample size. Tullis and Stetson (2004) investigated which standardized usability scale most quickly converged on the "correct" conclusion regarding the usability of two websites as a function of sample size–where "correct" conclusion meant a significant t-test consistent with the decision reached using the total sample size. The study showed that the SUS was the fastest to converge on the correct conclusion, reaching 100% agreement at a sample size of 12. But it is not known if the same is true for UMUX-Lite.

Finally, UMUX-Lite is a relatively new scale and its reliability, validity and sensitivity in different situations for different types of interfaces is still under study. For example, we know that Bangor et al. (2008) found the SUS to be sensitive to differences among types of interfaces and changes made to a product. At the moment it is not fully established to what extent UMUX-Lite is sensitive to differences among types of interfaces and especially for the chatbots. The current study found that it is reliable measurement usability in chatbots. Nonetheless, there uncertainty regarding the content validity of the UMUX-Lite when it is applied to chatbots. Is UMUX-Lite able to fully assess the usability of the chatbot? Are all relevant aspects included?

We can see that from 18 factors that participants found important, UMUX-lite only measures one factor (perceived ease of use). Thus, it might be assumed that UMUX-Lite can inform about the usability of a chatbot, but it does not include all relevant aspects. This might explain why in Toshiba pair despite the overall differences in perceived usability, the result of the t-test was not significant. During the debriefing, participants mentioned not liking Yoko's synthetic voice or the fact that instead of full answers she only presented more links. Nevertheless, they managed to finish the task and retrieved information they needed. Consequently, when users were administered UMUX-lite they only reported perceived ease of

use and usefulness. Information about the attitude towards other factors (voice tone, variation in responses, the maxim of relation and etc.) was not fully reflected in UMUX-Lite results. This might be the reason why despite reported quantitative and qualitative differences UMUX-Lite was not sensitive enough to produce a significant difference.

6.3 General Limitations

Several limitations may influence the interpretation and application of this study. Even though surveys presented descriptions for each factor, it is not clear how respondents interpreted the descriptions and the items of the survey. Additionally, although online survey and interaction test participants were given the same set of factors, descriptions of the statements were slightly different. This inconsistency could have affected the outcome of the results. Surveys used to measure the importance of the factors in the online survey and interaction test were not validated, therefore, there is a chance of error in results. Moreover, there is a chance that biases like task-selection, Hawthorne effect, social desirability, recency and primacy effects, could have affected the results of the research.

Furthermore, even though this survey attempted to capture most of the main factors, there is a possibility that some factors were overlooked. Analysis of the debrief showed that participants also find important good linguistic processing capabilities of the chatbot that allows them to be more flexible with their input (see Section 4.2.3 and 5). Similarly, some other factors could have been missed during the literature review or interaction test due to the small sample size.

As it was mentioned in the methods section, an Italian company supported this project in kind (UserBot.ai) and they helped in the diffusion of the survey. Therefore, about 95% of experts that participated in the online survey were Italian. The fact that the majority of experts were from UserBot.ai, Italian, make results from online survey less generalizable to the experts from other parts of the world.

6.4 Recommendations

Results from the current study with a modest sample size showed that UMUX-Lite, to some degree, is sensitive to differences and is a reliable measure of usability. However, there is a concern about the extent of the sensitivity and validity of the scale. Research showed that UMUX-lite does not include all the aspects important for measuring the perceived usability of the chatbot. This finding implies that there is a need for a post-test questionnaire that will have

higher content validity and will be able to capture more aspects of interaction with the chatbot. Additionally, there is a need for the questionnaire that will not only inform about satisfaction level of the users but will also help to explore strong and weak aspects in designed chatbots.

Based on the findings of the present study it is feasible to develop a questionnaire that will incorporate UMUX-LITE as a measure of satisfaction but will also include items to measure key factors associated with the interaction with chatbots (see Section 5). Such a questionnaire also has the advantage of being a relatively easy method to investigate perceived satisfaction of the users, as questionnaires are easier to administer and analyze (Jones, Murphy, Edwards, & James, 2008). Moreover, factors developed for the questionnaire can be used as a checklist for the developers. This might help developers to ensure that they are not missing essential factors in their design.

7. CONCLUSION

The goal of this exploratory study was to identify a list of key factors that shape the perception of usability in interaction with chatbots and propose direction for the development of the new tool based on the key factors and existing tools. This goal was achieved through a systematic literature review, an online survey and interaction test. While this study has focused solely on chatbots in information retrieval tasks, the findings have the potential to be applied to and built upon for other CAs in the future.

The systematic literature review showed that CIs have a different set of factors that influence perceived usability. From the 28 factors found from the literature and debriefing, 18 were perceived to be important by online survey and interaction test participants. Investigation of the UMUX-Lite as a measurement of usability of chatbots showed that even though UMUX-Lite is reliable and reasonably sensitive, it does not include all the aspects important in interaction with chatbots. To conclude, for further research, this study proposes to design a questionnaire that will be based on the key factors established in this study and will incorporate UMUX-Lite.

References

Amazon. (n.d.-a). Alexa Design Guide - Voice. Retrieved from https://developer.amazon.com/docs/alexa-design/voice-experience.html#guidelines-for-voice-interactions

Amazon. (n.d.-b). Alexa Design Guide - Voice Experiences. Retrieved from https://developer.amazon.com/docs/alexa-design/design-voice.html

Applin, S. A., & Fischer, M. D. (2015). New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. *2015 Ieee International Symposium on Technology and Society (Istas)*. https://doi.org/10.1109/istas.2015.7439436

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495. https://doi.org/10.1080/10447318.2015.1064648

Chakrabarti, C., & Luger, G. F. (2015). Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, *42*(20), 6878–6897. https://doi.org/10.1016/j.eswa.2015.04.067

Cohen, D., & Lane, I. (2016). An oral exam for measuring a dialog system's capabilities. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 835–841.

Coniam, D. (2014). *The linguistic accuracy of chatbots: Usability from an ESL perspective*. *Text & Talk* (Vol. 34). https://doi.org/10.1515/text-2014-0018

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews.

Knowledge in Society, *1*(1), 104–126. https://doi.org/10.1007/BF03177550

EndNote X8. (n.d.). EndNote X8. Boston, MA, USA. Retrieved from https://endnote.com/product-details/

Finstad, K. (2010). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/https://doi.org/10.1016/j.intcom.2010.04.004

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

Gnewuch, U., Morana, S., & Maedche, A. (2018). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *ICIS 2017: Transforming Society with Digital Innovation*. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041742966&partnerID=40&md5=e04f4be1dea36cccd52963bb8da7106f

Google. (2017). Conversation design– Learn about conversation. Retrieved from https://designguidelines.withgoogle.com/conversation/conversation-design/learn-about-conversation.html

Griol, D., Carbó, J., & Molina, J. M. (2013). An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence*, *27*(9), 759–780. https://doi.org/10.1080/08839514.2013.835230

Hertzum, M., Andersen, H. H. K., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers*, *14*(5), 575–599. https://doi.org/10.1016/s0953-5438(02)00023-1

Huang, H. Y., Li, Y. H., Lin, J. M., & Yang, D. L. (2015). A Study on Psychological Care for the Elderly Using Web-Based Embodied Conversational Agent. *Journal of Internet Technology*, *16*(1), 35–45. https://doi.org/10.6138/jit.2014.16.1.20130821

Johnson, K. (2018). Facebook Messenger passes 300,000 bots | VentureBeat. Retrieved January

7, 2019, from https://venturebeat.com/2018/05/01/facebook-messenger-passes-300000-bots/

Jones, S., Murphy, F., Edwards, M., & James, J. (2008). *Doing things differently: advantages and disadvantages of web questionnaires*. *Nurse Researcher*. Retrieved from https://s3.amazonaws.com/academia.edu.documents/44240997/Doing_things_differently_advantages_and_20160330-546-zw48tm.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1546864188&Signature=I2NHnq87uxkNIvZhL6j6SPlDQ8o%3D&response-content-disposition=inline%3B

Kirakowski, J., Odonnell, P., & Yiu, A. (2009). Establishing the Hallmarks of a Convincing Chatbot-Human Dialogue. In *Human-Computer Interaction*. https://doi.org/10.5772/7741

Kojouharov, S. (2018). How Businesses are Winning with Chatbots & Ai – Chatbots Life. Retrieved January 5, 2019, from https://chatbotslife.com/how-businesses-are-winning-with-chatbots-ai-5df2f6304f81

Kuligowska, K. (2015). *Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Conversational Agents* (Vol. 2). https://doi.org/10.18483/PCBR.22

Lee, S. Y., & Choi, J. H. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, *103*, 95–105. https://doi.org/10.1016/j.ijhcs.2017.02.005

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

McTear, M., Callejas, Z., Barres, D. G., & article, F. G. (2016). *The Conversational Interface - Talking to Smart Devices*. *Springer International Publishing* (1st ed.). Springer

International Publishing. https://doi.org/DOI 10.1007/978-3-319-32967-3

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

Meira, M. de O., & Canuto, A. M. de P. (2015). Evaluation of Emotional Agents' Architectures-an Approach Based on Quality Metrics and the Influence of Emotions on Users. In *World Congress on Engineering* (Vol. 1). London.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nguyen, M.-H. (2017). Chatbot Market 2017: Stats, Trends, Size & Ecosystem Research - Business Insider. Retrieved January 5, 2019, from https://www.businessinsider.com/chatbot-market-stats-trends-size-ecosystem-research-2017-10?international=true&r=US&IR=T

Pauletto, S., Balentine, B., Pidcock, C., Jones, K., Bottaci, L., Aretoulaki, M., … Balentine, J. (2013). Exploring expressivity and emotion with artificial voice and speech technologies. *Logopedics Phoniatrics Vocology*, *38*(3), 115–125. https://doi.org/10.3109/14015439.2013.810303

Peeters, M. M. M. (2016). *ReMindMe: Agent-based Support for Self-disclosure of Personal Memories in People with Alzheimer's Disease*. (C. Rocker, M. Ziefle, J. Odonoghue, L. Maciaszek, & W. Molloy, Eds.), *Proceedings of the International Conference on Information and Communication Technologies for Ageing Well and E-Health*. https://doi.org/10.5220/0005913000610066

Polisena, J., Castaldo, R., Ciani, O., Federici, C., Borsci, S., Ritrovato, M., … Pecchia, L. (2019).

Method HEALTH TECHNOLOGY ASSESSMENT METHODS GUIDELINES FOR MEDICAL DEVICES: HOW CAN WE ADDRESS THE GAPS? THE INTERNATIONAL FEDERATION OF MEDICAL AND BIOLOGICAL ENGINEERING PERSPECTIVE. *International Journal of Technology Assessment in Health Care, 34*(3), 276–289. https://doi.org/10.1017/S0266462318000314

Qualtrics. (2018). Qualtrics Research Core. Provo, Utah, USA: Qualtrics. Retrieved from https://www.qualtrics.com/research-core/

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

Radziwill, N. M., & Benton, M. C. (2017). *Neurodiversity secrets for innovation and design.* SXSW Interactive, Austin TX. Retrieved from https://schedule.sxsw.com/2017/events/PP66934

Ramos, R. (2017). Screw the Turing test — chatbots don't need to act human. Retrieved from https://venturebeat.com/2017/02/03/screw-the-turing-test-chatbots-dont-need-to-act-human/

Randolph, J. J. (2008). Online Kappa Calculator [Computer software]. Retrieved from http://justusrandolph.net/kappa/

Randolph, J. J. (2009). A Guide to Writing the Dissertation Literature Review, *14*. Retrieved from http://lemass.net/capstone/files/A Guide to Writing the Dissertation Literature Review.pdf

Sauro, J. (2011). *A practical guide to the system usability scale : background, benchmarks & best practices*. Denver: Measuring Usability LCC.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research*. Elsevier Inc.

Solomon, M. (2017). If Chatbots Win, Customers Lose, Says Zappos Customer Service Expert. Retrieved from https://www.forbes.com/sites/micahsolomon/2017/03/23/customers-lose-if-chatbots-win-says-zappos-customer-service-expert/#24c815576087

Staven, T. (2017). What Makes a Good Bot or Not? UNIT4. Retrieved from https://www.unit4.com/blog/2017/03/what-makes-a-good-bot-or-not

Toplin, J. (2017). *The Conversational Commerce Report*. *BI Intelligence*. Retrieved from https://www.businessinsider.com/intelligence/research-store?IR=T&utm_source=businessinsider&utm_medium=report_teaser&utm_term=report_teaser_store_text_link_chatbot-market-stats-trends-size-ecosystem-research-2017-10&utm_content=report_store_report_teaser_

Tubbs, S. L. (2013). *Human communication : principles and contexts*. McGraw-Hill.

Tullis, T. S., & Stetson, J. N. (2004). *A Comparison of Questionnaires for Assessing Website Usability*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.3677&rep=rep1&type=pdf

Van Eeuwen, M. (2017a). Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers. *University of Twente*, (essay:71706), 15. Retrieved from http://essay.utwente.nl/71706/1/van Eeuwen_MA_BMS.pdf%0Ahttp://essay.utwente.nl/71706/%0Ahttp://essay.utwente.nl/71706/1/van Eeuwen_MA_BMS.pdf

Van Eeuwen, M. (2017b). *Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers*. University of Twente. Retrieved from https://essay.utwente.nl/71706/1/van Eeuwen_MA_BMS.pdf

Vetter, M. (2002). Quality Aspects of Bots. In D. Meyerhoff, B. Laibarra, R. van der Pouw Kraan, & A. Wallet (Eds.), *Software Quality and Software Testing in Internet Times* (pp.

165–184). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56333-1_11

Wilson, H. J., Daugherty, P. R., & Morini-Bianzino, N. (2017). The Jobs That Artificial Intelligence Will Create. *MIT SLOAN MANAGEMENT REVIEW*, (Summer 2017). Retrieved from http://mitsmr.com/2odREFJ

Appendix A

The Systematic Literature Review Plan

The rationale of the literature review is to formulate a list of key factors that contribute to the quality of interaction with chatbots.

This literature review's contribution to existing research: This review is motivated by practical concerns and lack of information, at the moments there are no defined criteria for interaction with a chatbot.

The focus	Will be articles that include findings and theories of quality factors of chatbot and articles that will include assessment methods for interaction quality with chatbots.
The goal	Integrate and generalize previous findings and propose a list of the key factors that affect interaction with a chatbot.
Perspective	The language of the literature review will be neutral.
Coverage	The review will only cover central or pivotal literature.
Organization	The review will be organized around propositions in a research rationale.
Audience	*Primary*- Reviewers of the work (1st and 2nd supervisor). *Secondary*- Other scientists, experts that were included in the research.
Methodology	This literature review will be qualitative and will follow the Phenomenological Method of the literature review.
Inclusion criteria	• Studies that mention chatbots or conversational interfaces/agents in their Title, Abstract or Keywords. • Studies that include findings and theories on factors that can potentially contribute perceived usability of CA • Studies that include assessment methods that might inform about criteria used during the assessment. Database search:

	• Studies from past 10 years.
Exclusion criteria	• Documents that talked about technical aspects and measurements of the CA and documents that did not inform about CI's interactional characteristics. • Studies that was not able to contribute to the list of factors were excluded Database search • Studies that are about virtual assistants (such as Google Assistant, Cortana, Siri, Alexa).
Search Inquiry - Scopus	(TITLE-ABS-KEY (chatbot* OR "conversational agents*" OR "conversational interface*") AND TITLE-ABS-KEY (interact*) AND TITLE-ABS-KEY (satis* OR quali*) AND NOT TITLE-ABS-KEY ("virtual assistant" OR voice)) AND PUBYEAR > 2009 AND (LIMIT-TO (SRCTYPE , "p") OR LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar"))
Search Inquiry - Web of Science	You searched for: TOPIC: (Chatbot* OR "conversational agent*" OR "conversational interfaces") *AND*TOPIC: (Interact*) *AND* TOPIC: (Satisf* or qual*) Refined by: LANGUAGES: (ENGLISH OR PORTUGUESE) AND PUBLICATION YEARS: (2018 OR 2014 OR 2010 OR 2017 OR 2013 OR 2016 OR 2012 OR 2015 OR 2011) AND DOCUMENT TYPES: (PROCEEDINGS PAPER OR ARTICLE) Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.
Tools	Prisma Flow diagram, PRISMA 2009 Checklist.

Appendix B

Expert and End-Users survey: Assessment of Chatbots' Perceived Usability

---

Expert and End-Users survey: Assessment of Chatbots' Perceived Usability

Start of Block: Introduction

Info Information Sheet and consent formIntroduction

The University of Twente (NL) in collaboration with University of Perugia (IT) and with the support of UserBot (https://userbot.ai/) are investigating the factors that may affect **end-users** quality of interaction with chatbots and conversational agents. Gunay Tariverdiyeva (g.tariverdiyeva@student.utwente.nl) supervised by Dr Simone Borsci (s.borsci@utwente.nl) from the University of Twente are conducting this first phase of the study.

Purpose of the studyTo explore and model with experts designers and end-users key aspects, identified by a literature review, that play a key role in the interaction experience with a chatbot. This will help us to define a framework and to build and standardised an evaluation tool to support the rapid assessment of people experience with chatbots. Sections of the survey, time to complete and procedureThe survey will take you approximately **15 minutes for experts and about 10 minutes for end-users** to be completed.

Rights of participants

There are no right or wrong answers in this study, we are interested in your opinion and you have the right to express positive and negative comments.

Moreover, you have the right to quit the experiment at any time.

Risks and data management

We believe there are no known risks associated with this study; however, as with any online related activity, the risk of a breach of confidentiality is always possible. To the best of our ability your answers in this study will remain confidential and anonymized, and data will be secured and stored in an encrypted repository.

Use of data and GDPR

Anonymized data will be used for statistical and research purposes, and data analysis will be used for reports and scientific publications. Anonymized data will be stored in a safe memory unit with encrypted access at the University of Twente and under control of Dr Simone Borsci.

Contacts

Your participation in this study is completely voluntary and you can withdraw at any time. If you have any questions concerning your rights as a participant to this study, you may contact Dr Simone Borsci (s.borsci@utwente.nl) Please print a copy of this page for your records. Thank you!

Q1 1. I read the information sheet and I accept to participate in this study

- ○ Yes, I read and I consent to be part of this study (1)
- ○ No (2)

End of Block: Introduction

Start of Block: Block 5

Q2 2. When it comes to chatbots and conversational agents, you can describe your-self as

- Expert professional e.g., programmer, designer, scientists etc. (1)
- End User who usually interact with chatbots and conversational agents (2)
- Someone who has no idea what chatbots or conversational agents are? (3)

End of Block: Block 5

Start of Block: Participant information

*

Age P1. Your age

_____

Nation P2. Your nationality

_____

Gender P3. Your gender

- Man (4)
- Female (5)
- Prefer not to say (6)

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Expert_Profession PE1. Please briefly describe your profession in terms of main duties associated to chatbots

_____

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Expert_year of exper PE2. How many years of expertise do you have as an expert in the chatbots and conversational agents field

- ○ From 0 to 2 (1)
- ○ More than 2 less than 5 (2)
- ○ More than 5 (3)

Display This Question:
If Q2 = End User who usually interact with chatbots and conversational agents

User Education P4. What is your level of education?

- ○ Less than high school (15)
- ○ High school graduate (16)
- ○ Some college (17)
- ○ Bachelor degree (18)
- ○ Master Degree (19)
- ○ Professional degree (20)
- ○ Doctorate (21)

Display This Question:

If Q2 = End User who usually interact with chatbots and conversational agents

User_time of use P5. In the last 12 months how many time you used chatbots?

- ○ I didn't in the last 12 months but I was used to. I consider my self an intermediate/good user (4)
- ○ Once (5)
- ○ Once at month (6)
- ○ at least once per week (7)
- ○ More than once per week (8)
- ○ Every day (9)

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Expert_methods_other PE3. On the basis of your knowledge, in your field which are the most frequently used approaches applied by OTHER COMPANIES to test the end users perceived quality of the intercation with chatbot? (select all the appropriate)

- ☐ Expert evaluation only (4)
- ☐ Observation of users interaction (5)
- ☐ Interview (6)
- ☐ Usability test with small groups (7)
- ☐ Usability test with large group (8)
- ☐ Remote Log Analysis (9)
- ☐ Conversational Analysis (10)
- ☐ Other please specify (11) _____

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

experts_Methods_them PE4. Which approaches do YOU usually use to test the end users perceived quality of the intercation with chatbot? (select all the appropriate)

- ☐ Expert evaluation only (4)
- ☐ Observation of users interaction (5)
- ☐ Interview (6)
- ☐ Usability test with small groups (7)
- ☐ Usability test with large group (8)
- ☐ Remote Log Analysis (9)
- ☐ Conversational Analysis (10)
- ☐ Other please specify (11) _____

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q7 PE5. Please briefly list, what are the main aspects you take into account when you design chatbots?

_____

End of Block: Participant information

Start of Block: List of factors that affect perceived usability of the chatbot

EXP

In the next section, we will present to you a list of 27 aspects extracted from the literature that

can play an important role to **define the experience of an end-user (high satisfaction and high perceived usability) during the interaction with a chatbot.**

Each factor has a brief description. Please familiarize yourself with the factors and state to what extent you agree or disagree on the effect these factors have on the quality of interaction with the chatbot. While you read the factors please think of how the factors can be improved. Your comments will be gathered at the end of the survey.

Page Break

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

PrimingP1 To measure people satisfaction with a chatbot it is important to ask to end users about ...

Display This Question:
If Q2 = End User who usually interact with chatbots and conversational agents

Q69 As an end user of chatbots I believe that my satisfaction is strongly affected by the folowing aspects of a chatbot ...

F1 **...Time response**

The ability of the chatbot to respond timely to users' input and requests.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Timely Response (1)	○	○	○	○	○	○	○

F2 **...Multi-thread conversation**

The ability of the chatbot to recognize process multiple parallel themes simultaneously. Ex: Set meeting with Jeff for tomorrow at 2 pm and cancel meeting with Maria at 1:30 pm.

	Strongly disagree (8)	Disagree (9)	Somewhat disagree (10)	Neither agree nor disagree (11)	Somewhat agree (12)	Agree (13)	Strongly agree (14)
Multi-thread conversation (1)	○	○	○	○	○	○	○

F3 **...Maxim of quantity** The ability of the chatbot to make its response as informative as is required but at the same time do not make its contribution more informative than is

	Strongly disagree (8)	Disagree (9)	Somewhat disagree (10)	Neither agree nor disagree (11)	Somewhat agree (12)	Agree (13)	Strongly agree (14)
Maxim of quantity (1)	○	○	○	○	○	○	○

F4 **...Maxim of quality** The ability of the chatbot to avoid saying what it is believed to be false or say things for which it lack's adequate evidence.

	Strongly disagree (8)	Disagree (9)	Somewhat disagree (10)	Neither agree nor disagree (11)	Somewhat agree (12)	Agree (13)	Strongly agree (14)
Maxim of quality (1)	○	○	○	○	○	○	○

F5 **...Maxim of manners**

The ability of the chatbot to make it is purpose clear and communicate clearly, without obscurity or ambiguity.

	Strongly disagree (8)	Disagree (9)	Somewhat disagree (10)	Neither agree nor disagree (11)	Somewhat agree (12)	Agree (13)	Strongly agree (14)
Maxim of manners (1)	○	○	○	○	○	○	○

F6 **...Maxim of relation**

The ability of the chatbot to be relevant and to provide an appropriate contribution to immediate needs at each stage of the communication exchange.

	Strongly disagree (8)	Disagree (9)	Somewhat disagree (10)	Neither agree nor disagree (11)	Somewhat agree (12)	Agree (13)	Strongly agree (14)
Maxim of relation (1)	○	○	○	○	○	○	○

F7 **...Appropriate degrees of formality**

The ability of the chatbot to use language style that is appropriated to the context.

	Strongly disagree (15)	Disagree (16)	Somewhat disagree (17)	Neither agree nor disagree (18)	Somewhat agree (19)	Agree (20)	Strongly agree (21)
Appropriate degrees of formality (4)	○	○	○	○	○	○	○

F8 **...Reference to what is on the screen**

Chatbot should be able to make use of environment it is embedded in. Everything on the screen is part of the context of the conversation and can be used to guide the user towards its goal.

	Strongly disagree (15)	Disagree (16)	Somewhat disagree (17)	Neither agree nor disagree (18)	Somewhat agree (19)	Agree (20)	Strongly agree (21)
Reference to what is on the screen (4)	○	○	○	○	○	○	○

F9 **...Visual Look**

The outward appearance of a chatbot's dialog box and avatar (cartoon-like, video sequence depicting a live person, disembodied, static/animated etc.)

	Strongly disagree (36)	Disagree (37)	Somewhat disagree (38)	Neither agree nor disagree (39)	Somewhat agree (40)	Agree (41)	Strongly agree (42)
Visual look (5)	○	○	○	○	○	○	○

F10 **...Voice Tone**

The degree of chatbot's spoken expressiveness (inflection, emotional information through tone) and the accuracy of the text-to-speech function

	Strongly disagree (15)	Disagree (16)	Somewhat disagree (17)	Neither agree nor disagree (18)	Somewhat agree (19)	Agree (20)	Strongly agree (21)
Voice Tone (5)	○	○	○	○	○	○	○

Page Break

Display This Question:
If Q2 = End User who usually interact with chatbots and conversational agents

Q71 As an end user of chatbots I believe that my satisfaction is strongly affected by the folowing aspects of a chatbot ...

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

PrimingP2 To measure people satisfaction with a chatbot it is important to ask to end users about ...

F11 **...Integration with the website**

Chatbots form of implementation on the website and visibility can affect users perceived usability (all pages/specific pages, floating window/pull-out tab/permanent etc.)

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Integration with the website (1)	○	○	○	○	○	○	○

F12 **...Graceful responses in unexpected situations**

The ability of the chatbot to gracefully handle unexpected input, communication mismatch and broken line of conversation by clarifying previous statements, changing the topic of conversation or providing escalation channels to human operators etc.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Graceful responses in unexpected situations (1)	○	○	○	○	○	○	○

F13 **...Recognition and facilitation of users' goal and intent**

The ability of the chatbot to recognize user's intent and guide the user to its goal by refining offerings based on the recognized intent.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Recognition and facilitation of users' goal and intent (1)	○	○	○	○	○	○	○

F14 **...Variation of responses**

The ability of the chatbots to respond in various ways to similar requests. Users pay more attention when there's more variation in conversation.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Variation of responses (1)	○	○	○	○	○	○	○

F15 **...Perceived Ease Of Use**

The degree to which a person believes that using mobile messenger chatbots would be free of effort.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Perceived Ease Of Use (1)	○	○	○	○	○	○	○

F16 **...Engage in on-the-fly problem solving**

The ability of the chatbot to solve problems instantly on the spot.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Engage in on-the-fly problem solving (1)	○	○	○	○	○	○	○

F17 **...Themed discussion**

Ability (or lack thereof) to maintain a conversational theme once introduced and to keep track of the context to understand the user's utterances.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Themed discussion (1)	○	○	○	○	○	○	○

F18 **...Breadth of knowledge**

The ability of the chatbot to exhibit knowledge that it is out of its immediate domain by

recognizing social context and applying current social trends in course of conversation to make the conversation more relevant and dynamic.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Breadth of knowledge (1)	○	○	○	○	○	○	○

F19 **...Initiative**

The ability of the chatbot to initiate conversation or offer cues for further discussion by presenting its functionality, offering a range of topics for discussion, directing to hyperlinks, suggesting further actions, etc.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Initiative (1)	○	○	○	○	○	○	○

F20 **...Personality**

The ability of the chatbot to convey personality, warmth, and authenticity by providing

greetings, self-introductory information (age, sex, interests, opinions), self-disclosure, reciprocity (mutual exchange of information), and empathetical responses.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Personality (1)	○	○	○	○	○	○	○

Page Break

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

PrimingP3 To measure people satisfaction with a chatbot it is important to ask to end users about ...

Display This Question:
If Q2 = End User who usually interact with chatbots and conversational agents

Q70 As an end user of chatbots I believe that my satisfaction is strongly affected by the folowing aspects of a chatbot ...

F21 **...Interaction enjoyment**

The impression that a technological device is enjoyable to operate, regardless of whether it provides. In comparison to regular interface CI can be more interactive it can make a joke, insert emojis or gif, etc.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Interaction enjoyment (1)	○	○	○	○	○	○	○

F22 **...Read and respond to moods of human participant**

The ability of the chatbot to appropriately recognize mood of the user from its utterances and respond accordingly.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Read and respond to moods of human participant (1)	○	○	○	○	○	○	○

F23 **...Read and respond to moods of human participant**

The ability of the chatbot to appropriately recognize mood of the user from its utterances and respond accordingly.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Read and respond to moods of human participant (1)	○	○	○	○	○	○	○

F24 **...Sensitivity to safety and social concerns**

The ability of the chatbot to recognize, respond and refer a user to helpline if signs of safety or social concern present in the conversation e. g. users mention the intent to commit a bad action

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Users' privacy and ethical decision making (1)	○	○	○	○	○	○	○

F25 **...Meets neurodiverse needs**

Chatbot meets neurodiverse needs of different users whether it is healthy users, elderly users or users with dyslexia, autism, old, etc.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Meets neurodiverse needs (1)	○	○	○	○	○	○	○

F26 **...Trustworthiness**

The ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage with the chatbot

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Trustworthiness (4)	○	○	○	○	○	○	○

F27 **...Process facilitation and follow up**

The ability of the chatbot to facilitate the initiated process and follow up with the process.

	Strongly disagree (22)	Disagree (23)	Somewhat disagree (24)	Neither agree nor disagree (25)	Somewhat agree (26)	Agree (27)	Strongly agree (28)
Process facilitation and follow up (4)	○	○	○	○	○	○	○

End of Block: List of factors that affect perceived usability of the chatbot

Start of Block: Factors sorting

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q8 Please sort each one of the following 10 aspects below in one (and only one) of the following three groups:

Marginal Aspects. These aspects have little or none effect on user satisfaction (perceived quality of interaction) therefore it is of marginal importance to measure these aspects with the end-users. These aspects are the ones (important or less important) without which people may "survive" and still achieve their goal in a satisfactory way.

Core Aspects. It is extremely important to measure these aspects with end-users because these determine a good quality of interaction with any type of chatbot independently from context and modality of interaction with a chatbot**.** These aspects may directly affect people experience during and after the use of a chatbot.

Dependent Aspects. It could be important to measure these aspects that account for the diversity of the chatbots. These aspects may affect (somehow) people quality of interaction but are dependent to the context and modality of use e.g., type of chatbot, type of device used (desktop, mobile) type of platform (website, facebook) etc.

Marginal Aspects	Core Aspects	Dependent Aspects
_____ Time response (ability to respond timely to users' requests) (35)	_____ Time response (ability to respond timely to users' requests) (35)	_____ Time response (ability to respond timely to users' requests) (35)
_____ Multi-thread conversation (ability to recognize and process multiple parallel topics simultaneously). (36)	_____ Multi-thread conversation (ability to recognize and process multiple parallel topics simultaneously). (36)	_____ Multi-thread conversation (ability to recognize and process multiple parallel topics simultaneously). (36)
_____ Maxim of quantity (ability to respond in an informative way with adding	_____ Maxim of quantity (ability to respond in an informative way with adding	_____ Maxim of quantity (ability to respond in an informative way with adding

too information) (37)

_____ Maxim of quality (ability to avoid false statements/information) (38)

_____ Maxim of manners (ability to make it is purpose clear and communicate without ambiguity. (39)

_____ Maxim of relation (ability to provide relevant and appropriate contribution to people needs at each stage) (40)

_____ Appropriate degrees of formality (ability of the chatbot to use appropriate language style for the context) (41)

_____ Reference to what is on the screen (ability to use the environment it is embedded in to guide the user towards its goal. (42)

_____ Visual Look (The outward appearance of a chatbot's dialog box and/or avatar) (43)

_____ Voice Tone (Spoken expressiveness (inflection, emotional information through tone) and the accuracy of the text-to-speech function (44)

too information) (37)

_____ Maxim of quality (ability to avoid false statements/information) (38)

_____ Maxim of manners (ability to make it is purpose clear and communicate without ambiguity. (39)

_____ Maxim of relation (ability to provide relevant and appropriate contribution to people needs at each stage) (40)

_____ Appropriate degrees of formality (ability of the chatbot to use appropriate language style for the context) (41)

_____ Reference to what is on the screen (ability to use the environment it is embedded in to guide the user towards its goal. (42)

_____ Visual Look (The outward appearance of a chatbot's dialog box and/or avatar) (43)

_____ Voice Tone (Spoken expressiveness (inflection, emotional information through tone) and the accuracy of the text-to-speech function (44)

too information) (37)

_____ Maxim of quality (ability to avoid false statements/information) (38)

_____ Maxim of manners (ability to make it is purpose clear and communicate without ambiguity. (39)

_____ Maxim of relation (ability to provide relevant and appropriate contribution to people needs at each stage) (40)

_____ Appropriate degrees of formality (ability of the chatbot to use appropriate language style for the context) (41)

_____ Reference to what is on the screen (ability to use the environment it is embedded in to guide the user towards its goal. (42)

_____ Visual Look (The outward appearance of a chatbot's dialog box and/or avatar) (43)

_____ Voice Tone (Spoken expressiveness (inflection, emotional information through tone) and the accuracy of the text-to-speech function (44)

Q72

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q64 Please sort each one of the following 10 aspects below in one (and only one) of the following three groups:

Marginal Aspects. These aspects have little or none effect on user satisfaction (perceived quality of interaction), therefore it is of marginal importance to measure these aspects with the end-users. These aspects are the ones (important or less important) without which people may "survive" and still achieve their goal in a satisfactory way.

Core Aspects. It is extremely important to measure these aspects with end-users because these determine a good quality of interaction with any type of chatbot independently from context and modality of interaction with a chatbot. These aspects may directly affect people experience during and after the use of a chatbot.

Dependent Aspects. It could be important to measure these aspects that account for the diversity of the chatbots. These aspects may affect (somehow) people quality of interaction but are dependent to the context and modality of use e.g., type of chatbot, type of device used (desktop, mobile) type of platform (website, facebook) etc.

Marginal Aspects	Core Aspects	Dependent Aspects
_____ Integration with the website (position in the website and visibility (all pages/specific pages, floating window/pull-out tab/permanent etc.) (249)	_____ Integration with the website (position in the website and visibility (all pages/specific pages, floating window/pull-out tab/permanent etc.) (249)	_____ Integration with the website (position in the website and visibility (all pages/specific pages, floating window/pull-out tab/permanent etc.) (249)
_____ Graceful responses in unexpected situations (Ability to gracefully handle unexpected input, communication mismatch and	_____ Graceful responses in unexpected situations (Ability to gracefully handle unexpected input, communication mismatch and	_____ Graceful responses in unexpected situations (Ability to gracefully handle unexpected input, communication mismatch and

broken line of conversation (250)

_____ Recognition and facilitation of users' goal and intent (Ability to recognize user's intent and guide the user to its goal). (251)

_____ Variation of responses (Ability to respond in various ways to similar requests (252)

_____ Perceived Ease Of Use (The degree to which a person believes that to interact with a chatbot would be free of effort). (253)

_____ Engage in on-the-fly problem solving (Ability of the chatbot to solve problems instantly on the spot) (254)

_____ Themed discussion (Ability to maintain a conversational theme once introduced and to keep track of the context to understand the user's utterances). (255)

_____ Breadth of knowledge (ability to exhibit knowledge that it is out of its immediate domain during a conversation). (256)

_____ Initiative (The ability to initiate conversation (or offer cues) for further discussion by presenting its functionality, offering suggestions etc.) (257)

_____ Personality (Ability to convey personality, warmth, and authenticity by providing greetings, self-introductory, empathy, information etc. (258)

broken line of conversation (250)

_____ Recognition and facilitation of users' goal and intent (Ability to recognize user's intent and guide the user to its goal). (251)

_____ Variation of responses (Ability to respond in various ways to similar requests (252)

_____ Perceived Ease Of Use (The degree to which a person believes that to interact with a chatbot would be free of effort). (253)

_____ Engage in on-the-fly problem solving (Ability of the chatbot to solve problems instantly on the spot) (254)

_____ Themed discussion (Ability to maintain a conversational theme once introduced and to keep track of the context to understand the user's utterances). (255)

_____ Breadth of knowledge (ability to exhibit knowledge that it is out of its immediate domain during a conversation). (256)

_____ Initiative (The ability to initiate conversation (or offer cues) for further discussion by presenting its functionality, offering suggestions etc.) (257)

_____ Personality (Ability to convey personality, warmth, and authenticity by providing greetings, self-introductory, empathy, information etc. (258)

broken line of conversation (250)

_____ Recognition and facilitation of users' goal and intent (Ability to recognize user's intent and guide the user to its goal). (251)

_____ Variation of responses (Ability to respond in various ways to similar requests (252)

_____ Perceived Ease Of Use (The degree to which a person believes that to interact with a chatbot would be free of effort). (253)

_____ Engage in on-the-fly problem solving (Ability of the chatbot to solve problems instantly on the spot) (254)

_____ Themed discussion (Ability to maintain a conversational theme once introduced and to keep track of the context to understand the user's utterances). (255)

_____ Breadth of knowledge (ability to exhibit knowledge that it is out of its immediate domain during a conversation). (256)

_____ Initiative (The ability to initiate conversation (or offer cues) for further discussion by presenting its functionality, offering suggestions etc.) (257)

_____ Personality (Ability to convey personality, warmth, and authenticity by providing greetings, self-introductory, empathy, information etc. (258)

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q66 Please sort each one of the following 7 aspects below in one (and only one) of the following three groups:

Marginal Aspects. These aspects have little or none effect on user satisfaction (perceived quality of interaction), therefore it is of marginal importance to measure these aspects with the end-users. These aspects are the ones (important or less important) without which people may "survive" and still achieve their goal in a satisfactory way.

Core Aspects. It is extremely important to measure these aspects with end-users because these determine a good quality of interaction with any type of chatbot independently from context and modality of interaction with a chatbot. These aspects may directly affect people experience during and after the use of a chatbot.

Dependent Aspects. It could be important to measure these aspects that account for the diversity of the chatbots. These aspects may affect (somehow) people quality of interaction but are dependent to the context and modality of use e.g., type of chatbot, type of device used (desktop, mobile) type of platform (website, facebook) etc.

Marginal Aspects	Core Aspects	Dependent Aspects
_____ Interaction enjoyment (enjoyable and egaging to operate regardless of whether it provides) (72)	_____ Interaction enjoyment (enjoyable and egaging to operate regardless of whether it provides) (72)	_____ Interaction enjoyment (enjoyable and egaging to operate regardless of whether it provides) (72)
_____ Read and respond to moods of human participant (ability to appropriately recognize mood of the user from its utterances and respond accordingly) (73)	_____ Read and respond to moods of human participant (ability to appropriately recognize mood of the user from its utterances and respond accordingly) (73)	_____ Read and respond to moods of human participant (ability to appropriately recognize mood of the user from its utterances and respond accordingly) (73)

_____ Sensitivity to safety and social concerns (ability of the chatbot to recognize, respond and refer a user to helpline if needed (74)	_____ Sensitivity to safety and social concerns (ability of the chatbot to recognize, respond and refer a user to helpline if needed (74)	_____ Sensitivity to safety and social concerns (ability of the chatbot to recognize, respond and refer a user to helpline if needed (74)
_____ Meets diversity needs (abilty to meets needs of users independently form their helath, well-being, age etc. (75)	_____ Meets diversity needs (abilty to meets needs of users independently form their helath, well-being, age etc. (75)	_____ Meets diversity needs (abilty to meets needs of users independently form their helath, well-being, age etc. (75)
_____ Trustworthiness (ability to convey accountability and trustworthiness to increase willingness to engage) (76)	_____ Trustworthiness (ability to convey accountability and trustworthiness to increase willingness to engage) (76)	_____ Trustworthiness (ability to convey accountability and trustworthiness to increase willingness to engage) (76)
_____ Process facilitation and follow up (Ability to facilitate processes and to perform a follow up) (77)	_____ Process facilitation and follow up (Ability to facilitate processes and to perform a follow up) (77)	_____ Process facilitation and follow up (Ability to facilitate processes and to perform a follow up) (77)

End of Block: Factors sorting

Start of Block: Concluding questions

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q9 **C1. As an expert do you find this list to be relevant to the chatbots' perceived usability?**

_____

Display This Question:

If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q11 **C2. Are there any factors that you think should be modified or improved? If yes, please state factors that you think should be improved and how?**

_____

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q12 **C3. How do you think these factors can be incorporated in the assessment of chatbots percieved usability?**

_____

Display This Question:
If Q2 = Expert professional e.g., programmer, designer, scientists etc.

Q13 **C4. Use text-box below for additional comments.**

_____

*

Q54 **C5. Your Email! We may like to involve you in future phases of this research, please leave your email if you wish to help us in future.**

_____

End of Block: Concluding questions

Appendix C

Interaction Test Plan

Interaction Test Plan

1. Pre-Interaction Test

1.1 Greeting

<<<

Hi,

Welcome, take a seat..." small talk".

Let me shortly describe you what is this study about: This study is focused on exploring what makes chatbots user-friendly and how designers can measure the perceived quality of interaction with chatbots.

During the experiment, you will be given a number of tasks that you will be able asked to accomplish both via chatbot and web-navigation. You and your actions on screen will be recorded via usability testing software(show them how it will look like).

Do you know what chatbots are?

If YES: Short definition just to make sure user does not confuse chatbots with something else.

If NO: Give short definition and give examples.

>>>

1.2 Consent Form

Consent Form for Assessment of Chatbots' Perceived Usability in Information Retrieval Tasks

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 been verbally provided the study information dated _____. 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 involves video-recorded usability test, survey questionnaire completed by me, and the video recorded interview. I was informed that data collected from me will be destroyed at the end of the research.	☐	☐
Use of the information in the study		
I understand that the information I provide will be used in the graduation project and might be published in scientific publications.	☐	☐
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.	☐	☐
Optional: I agree that my information can be quoted in research outputs without disclosing my identity.	☐	☐
Future use and reuse of the information by others		
I give permission for the video recording and survey data that I provide to be archived in the Faculty of Behavioural, Management and Social sciences (BMS) of University of Twente so it can be used for future research on chatbots by Dr Simone Borsci. Data will be used only for the research that is carried out at BMS.	☐	☐

Signatures

Name of participant Signature Date

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

Gunay Tariverdiyeva

Researcher name Signature Date

UNIVERSITY OF TWENTE.

Study contact details for further information: For further details on this study you can contact Gunay Tariverdiyeva (g.tariverdiyeva@student.utwente.nl) or Dr Simone Borsci (s.borsci@utwente.nl).

Contact Information for Questions about Your Rights as a Research Participant

If you have questions or concerns about this study that you want to discuss with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-bms@utwente.nl

1.3 Test Flow

<<<

This experiment has 3 main parts:

- Pre-test survey- where you will be asked to fill in basic information about yourself, like age, nationality, how well you are familiar with chatbots
- Interaction test- is the main part of the experiment where you will be asked to perform short tasks with chatbot and web navigation.
- Post-test interview & survey- after your interactions with chatbots during the test I will ask you a couple questions about your experience and then I will ask you to fill in survey which will also help me to better understand your experience with chatbots.

>>>

1.4 Pre-test Survey

Intro
Pre Interaction Test Questionnaire
Please read carefully questions below and answer them.

Q1 Participant Code. This will be filled by the researcher.

_____ *

Q3 Age

_____

Q4 Nationality

_____

Q5 Gender

- ○ Male
- ○ Female
- ○ Other _____

Q6 Education Level

- ○ Less than high school
- ○ High school graduate
- ○ Bachelor's Degree
- ○ Master's Degree
- ○ Professional degree
- ○ Doctorate

Q7 English Language Proficiency

- native speaker
- near native (C2)
- excellent command (C1)
- very good command (B2)
- good command (B1)
- basic communication skills (A1, A2)

Q8 Employement

- Employed full time
- Employed part time
- Unemployed looking for work
- Unemployed not looking for work
- Retired
- Student
- Disabled

Q9 Occupation (Field of Study)

_____

Q10 How well informed or proficient you are in use of modern technology?

	Extremely well				Not well at all
How well informed or proficient you are in use of modern technology?	○	○	○	○	○

Q11 How familiar you are with chatbots or conversational interfaces?

	Extremely familiar	Very familiar	Moderately familiar	Slightly familiar	Not familiar at all
How familiar you are with chatbots or conversational interfaces?	○	○	○	○	○

Q12 Have you used chatbot or conversational interface before?

- ○ Definitely yes
- ○ Probably yes
- ○ Might or might not
- ○ Probably not
- ○ Definitely not

Display This Question:

If Have you used chatbot or conversational interface before? = Definitely yes

Or Have you used chatbot or conversational interface before? = Probably yes

Or Have you used chatbot or conversational interface before? = Might or might not

Q13 How often do you use it?

- Daily
- 4-6 times a week
- 2-3 times a week
- Once a week
- Rarely
- Never

Display This Question:

If Have you used chatbot or conversational interface before? = Definitely yes
Or Have you used chatbot or conversational interface before? = Probably yes
Or Have you used chatbot or conversational interface before? = Might or might not

Q14 What did you use chatbot or conversational interface for?

_____

Display This Question:

If Have you used chatbot or conversational interface before? = Definitely yes
Or Have you used chatbot or conversational interface before? = Probably yes
Or Have you used chatbot or conversational interface before? = Might or might not

2. Interaction Test

2.1 Interaction Test Script

<<< There will be 6 pairs of short tasks, 6 with a chatbot & 6 with a website. They will be presented to you in form of task cards. After each task, you will fill in short 2 questions questionnaire called UMUX-Lite. Each task card has "task code" written on it. You should write these task code on top of UMUX.

Each task will consist of a short realistic scenario. You as a participant should play along with this scenarios and imagine yourself in those situations. And when

encountered with the same task imagine that you are looking for that information for the first time.

While you are performing this scenario I would like to ask you to verbalize out loud each step you are doing and why, any issue you are experiencing in performing the task, or about the aspects you do not understand about the interface, the scenario and any technical problem you are facing during this test.

Now I will show you an example of how you should verbalize your thoughts so you know what is expected of you.

If I will notice that you will not verbalize I will try to remind you to do that. It could happen that I will ask you questions sometimes to better understand what you are doing and if you are experiencing an issue.

During the test, I will be in a neutral position and will seat at the back. You should perform tasks independently. However, if there something that you don't understand in the task description please let me know.

Once you feel that you have achieved the task of the scenario, or if you feel that the task is not achievable please report that to me. After that please fill in the UMUX-lite by writing down your task code.

I would like to emphasize that there is no wrong or right answer in this test. Each participant has their own way of navigating through the internet. Therefore, I would

appreciate if you could speak up your thoughts as it will help me to understand what users really think about chatbots and their behavior >>>

Ask if the participant has any question?

Ask if the participant is ready to start the test?

*** START RECORDING

2.2 Familiarization Task

Recheck the test set-up to make sure everything is ready for the test.

<<< Before we start main task tasks I want you to try to familiarize yourself with the chatbots and practice thinking aloud.

Please use chatbot Mitsuku to converse. Unlike other chatbots, you will talk to later Mitsuku is developed to converse on general topics so feel free to start the conversation about anything you want. You can find Mitsuku at m.me/chatbots.io. >>>

2.3 Main Tasks

<<< Now we will start the main part, as we already discussed earlier you will be given tasks that you have to complete either with chatbot or through regular web navigation. If you don't understand question please let me know, I'll try to clarify it. Once you start the task you will need to complete the task independently, it is ok if you can't complete the task it is also quite normal and expected since technology is not always designed clearly and can be confusing. Just let me know if you feel like you can complete the task on your own.

Also, as mentioned earlier let me know if you are finished with the task and ready to fill in UMUX-Lite.

This is your first task. Please read it and start acting on it. And please don't forget to verbalize your actions>>>

<<<

H1W

ou are living in New York, and planning vacation within the USA. Your travel dates are between 1st and 5th May. You have a budget limit of 500$ for flight tickets and hotel. Use hipmunk.com website to explore travel opportunities within your budget.

>>>

Now that you have completed the task please fill in UMUX-Lite.

Task Code: H1W

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
Chatbot's capabilities meet my requirements.	•	•	•	•	•		•
Chatbot is easy to use.	•	•	•	•	•		•

If you are ready we will continue to the next task. Present the next task.

Same sequence for the next 11 tasks……..

H2C

You are living in New York, and planning vacation within the USA. Your travel dates are between 1st and 5th May. You have a budget limit of 500$ for flight tickets and hotel. Use chatbot (m.me/hipmunk) to explore possible travel opportunities within your dates and budget.

F3W

You are moving from the Netherlands to Finland. You booked your tickets through Finnair. One of the things you want to take with you is your bicycle. But before you decide to take your bicycle to Finland you want to know what are the rules and costs of transporting a bicycle. Use https://www.finnair.com/int/gb/ website to find out how to transport a bicycle.

F4C

You are moving from the Netherlands to Finland. You booked your tickets through Finnair. One of the things you want to take with you is your bicycle. But before you decide to take your bicycle to Finland you want to know what are the rules and costs of transporting a bicycle. Use chatbot (m.me/Finnair) to find out how to transport a bicycle.

T5W

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. Use google.com website to find out how you can partition your hard drive?

T6C

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. Use chatbot (http://www.toshiba.co.uk/generic/yoko-home/) to find a solution.

I7W

You have an interview with Inbenta and you want to learn what is the address of Inbenta's Mexico office. Use Inbenta's website (https://www.inbenta.com/en/) to find the answer to your question.

I8C

You have an interview with Inbenta and you want to learn what is the address of Inbenta's Mexico office. Use Inbenta's chatbot on their website (https://www.inbenta.com/en/) to find the answer to your question.

A9W

You are on a trip to the USA and you are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet with an old friend for a couple of hours and see the city. Use https://www.amtrak.com/home website to find out the cost for temporarily storing your luggage at the station.

A10C

You are on a trip to the USA and you are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet with an old friend for a couple of hours and see the city. Find chatbot on https://www.amtrak.com/home to find out cost for temporarily storing your luggage at the station.

TA11W

You moved from the Netherlands to Australia recently. You want to know when is the deadline to lodge/submit your tax return. Use https://www.ato.gov.au/ website to find out when is the deadline?

TA12C

You moved from the Netherlands to Australia recently. You want to know when is the deadline to lodge/submit your tax return. Find chatbot on https://www.ato.gov.au/ to find out when is the deadline?

2.4 Task Sequence

Tasks will be presented on the cards(see below). Task cards will be organized in 2 decks according to the conditions(web/chatbot) and will be shuffled. The order will be

random as cards will be shuffled. Additionally, the start condition will also be randomized for the users. The user will be presented with one card at a time from one of the decks. In case the participant chooses a pair of the task participant just completed deck will be shuffled and the participant will be given next card.

Task Code: H1W

You are living in New York, and planning vacation within USA. Your travel dates are between 1st and 5th May.You have a budget limit of 500$ for flight tickets and hotel. Use hipmunk.com ***website*** to explore travel opportunities within your budget.

3. Post-Interaction Test

3.1 Interview

After the participant finished the last task and filled in UMUX-Lite participant will be interviewed. This will help to get a better understanding of user preferences when it comes to chatbots.

<<<

Thank you have done a great job if you are ready we will move to the next step?

Now, I will ask you questions about your experience with chatbots today. I will appreciate if you could give me your honest opinion and share your feelings about your experience as much as possible.

>>>

Based on the experience you had during the experiment:

1. What are the things that you didn't like during your interaction with chatbot?
2. What are the things that you liked during your interaction with chatbot?
3. Which chatbot did you enjoy interacting with the most? Why?
4. Which chatbot did you not enjoy interacting with? Why?
5. Can you tell me if you prefer more to us chatbots or website navigation? Why?
6. In which situations you would prefer using chatbots to the standard navigation?
7. In your opinion, what qualities or features were missing?
8. In your opinion, what qualities or features could be improved or added?

*** Remind participant that answers should be based on interaction participant had during the experiment.

3.2 Factors survey

Thanks for your answers, now we arrived at the last part. Please fill in this survey.

Post Test Questionnaire: Factors

Post Test Questionnaire: Factors

Start of Block: Factor Statements

Post-test Questionnaire: Factors
Below you will find below a list of statements that describe some of the characteristics of the chatbots. Please read them carefully, in case you encounter a statement that is not clear, please inform me. I'll try to clarify it for you and improve statement for future participants.

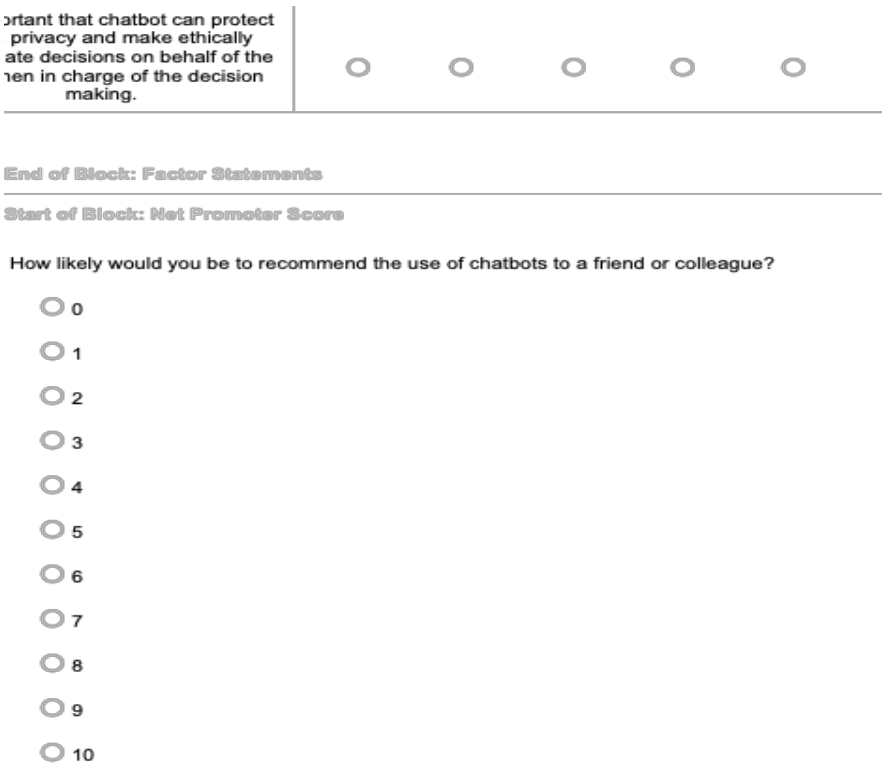
Q33 Participant Code

_____

ch do you agree or disagree with each statement in terms of importance to have a good ve experience with a chatbot?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agr
t chatbot to give me timely ponses- when I expect it.	○	○	○	○	○	
atbot to recognize and perform e requests at the same time.	○	○	○	○	○	
atbot to be informative without ing too much information.	○	○	○	○	○	
ortant that chatbot only gives nformation that is true.	○	○	○	○	○	
chatbot to communicate its ose and avoid ambiguity.	○	○	○	○	○	
chatbot to assist me in a way nsistent with my goals at each given stage.	○	○	○	○	○	
t chatbot to communicate in age style appropriate to the domain it is used for.	○	○	○	○	○	
atbot to reference other parts creen when it is necessary to e me with more information.	○	○	○	○	○	
features of the chatbot like a x and avatar are important for me.	○	○	○	○	○	
has a text-to-speech function, have an expressive voice and be accurate.	○	○	○	○	○	
hatbot to be easily accessible le on the page it is integrated, el of integration in a website.	○	○	○	○	○	
should not let discussion come nd even in situations when it understand or misinterprets user input.	○	○	○	○	○	

mportant that chatbot can ze what I want to accomplish nake suitable suggestions.	○	○	○	○	○
ortant that chatbot can vary its onses when asked similar ons instead of repeating the same answer.	○	○	○	○	○
ect chatbot to be easy and effortless to use.	○	○	○	○	○
with chatbot I can solve some asks right on the spot.	○	○	○	○	○
ortant that chatbot can keep ny previous inputs and pick up conversation where I left it.	○	○	○	○	○
nportant for chatbot exhibit ledge outside of it domain.	○	○	○	○	○
ortant for a chatbot to initiate ation and offer cues for further conversations.	○	○	○	○	○
ortant for a chatbot to convey ty, warmth, and authenticity by g greetings, self-introductory ormation, empathy, etc.	○	○	○	○	○
ortant to enjoy the interaction with a chatbot.	○	○	○	○	○
rtant for a chatbot to recognize spond suitably to my mood stration, happiness, etc.).	○	○	○	○	○
rtant for a chatbot to recognize pond to safety concerns and a user to helpline if needed.	○	○	○	○	○
portant for a chatbot to meet of users independent of their ations imposed by health itions, well-being, age, etc.	○	○	○	○	○
ortant for a chatbot to convey tability and trustworthiness to ase willingness to engage.	○	○	○	○	○
ortant for a chatbot to inform ite about the status of ongoing task.	○	○	○	○	○



ɔrtant that chatbot can protect privacy and make ethically ate decisions on behalf of the ıen in charge of the decision making.	○	○	○	○	○

End of Block: Factor Statements

Start of Block: Net Promoter Score

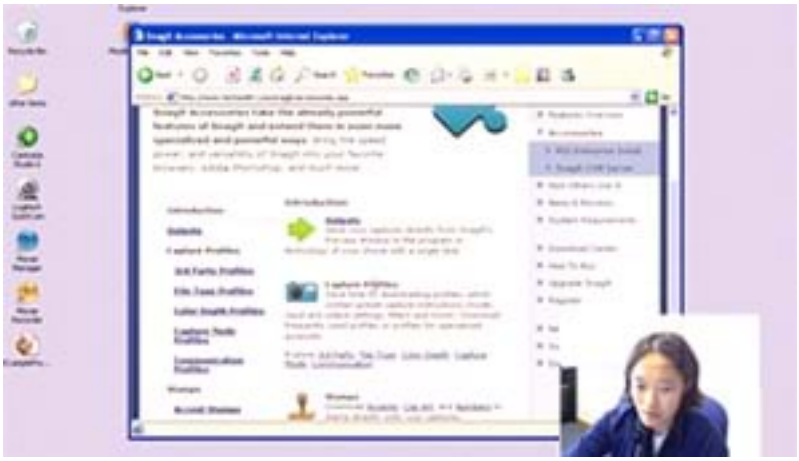
How likely would you be to recommend the use of chatbots to a friend or colleague?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10

4. Technical Details & Set-up

The test is conducted in a quiet room. Only people present in the room are a participant and test administrator.

Users will perform tasks using a computer and Google Chrome browser. The test will be recorded via usability software Morae and a webcam. More records screen, mouse movement, participant(webcam) and sound(webcam).

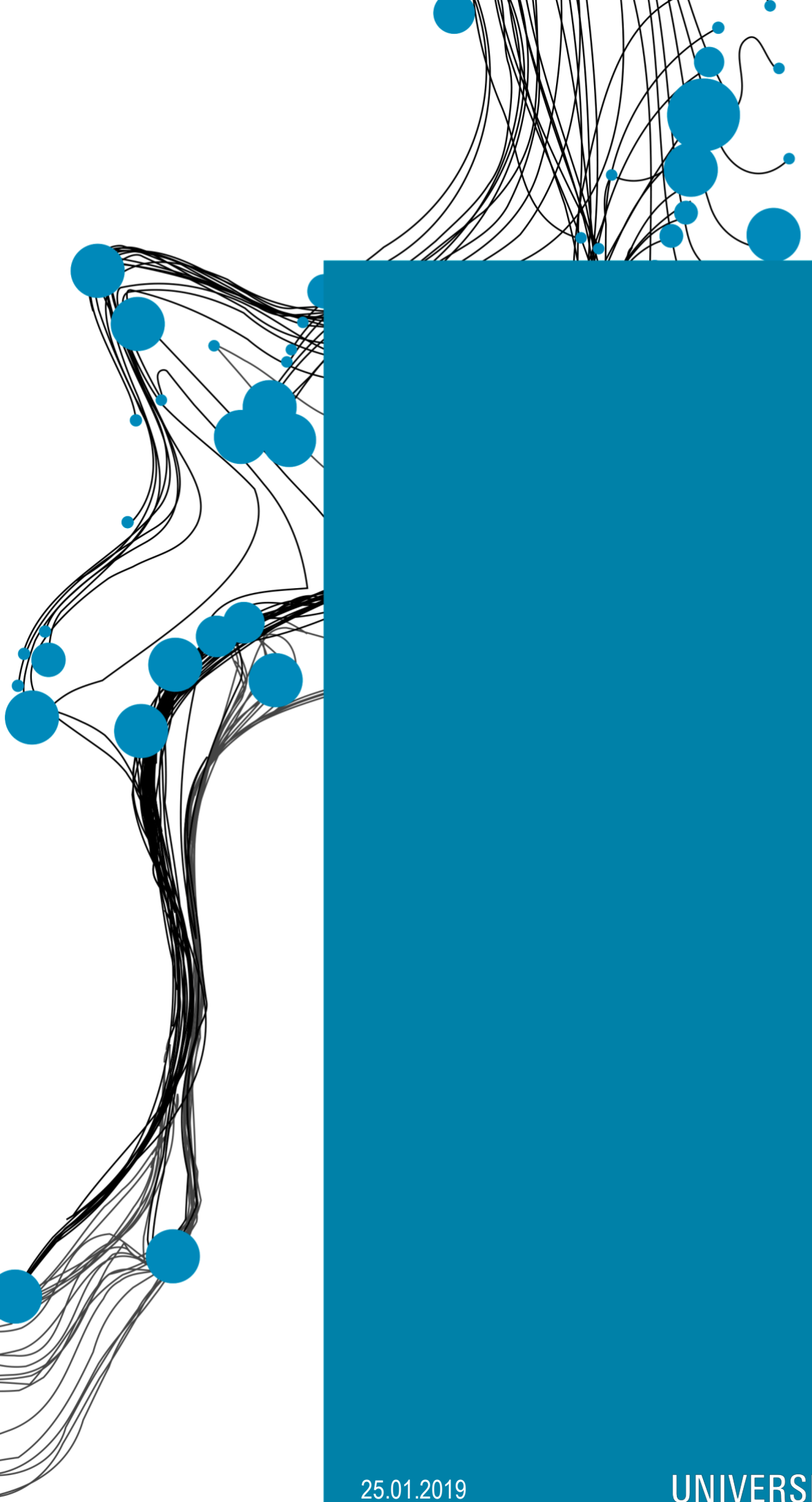


All surveys are presented and filled online with the help of Qualtrics software.

5. Interaction Test Checklist

- Greet Participant
- Small Talk
- Shortly describe study
- Ask if the participant is familiar with chatbots.
- Verbal Explanation of study information and data collection
- Consent form
- Pre-test survey
- Explain Interaction test in detail:
    - You will perform tasks
    - Task card
    - Chatbot vs Web
    - Task code/BOOKMARK
    - UMUX-Lite
    - How to act
    - Don't fill in everything
    - Think aloud

    - Think aloud example
    - I will be neutral
    - Task end
    - No right or wrong
- START RECORDING
- Interaction test
- Present tasks
- Interview
- STOP RECORDING
- Post Test Survey
- Feedback on the experiment and candy bar



25.01.2019

UNIVERSITY OF TWENTE.