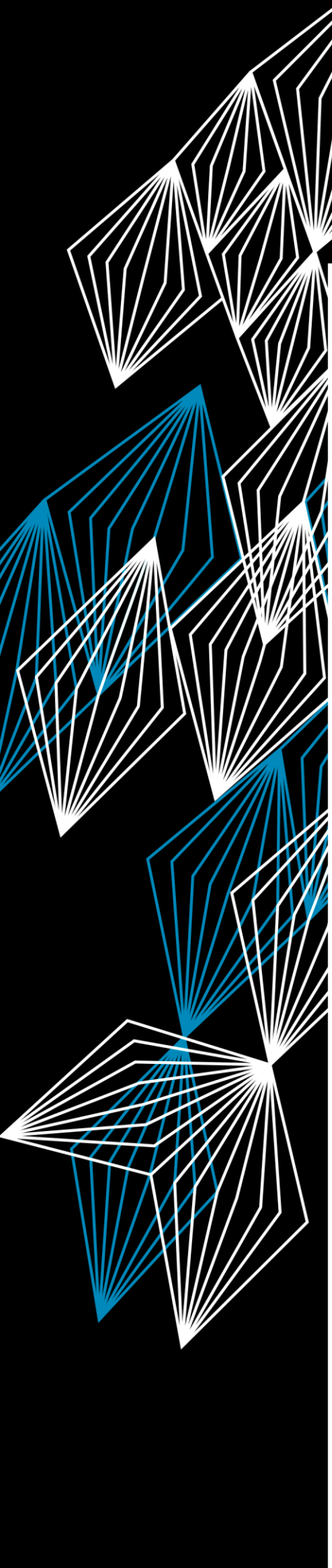


MASTER THESIS

Validity and Reliability of the User Satisfaction with Information Chatbots Scale (USIC)

Imke Silderhuis
September, 2020

Faculty of Behavioural, Management and Social Sciences (BMS)
Human Factors and Engineering Psychology

EXAMINATION COMMITTEE
Dr. S. Borsci
Dr. R. van der Lubbe

UNIVERSITY OF TWENTE.

Abstract

Although the chatbot market is growing, chatbots have difficulty to live up to their potential and often disappear due to disappointing usage (Brandtzaeg & Følstad, 2018). Developers need insight into which chatbot aspects users are satisfied with and which aspects need further improvement to retain their success. As of yet, there are no standardized scales available to assess the user's satisfaction with chatbots (Balaji & Borsci, 2019).

In the current study, we evaluated a promising scale that assesses user satisfaction with information chatbots (USIC). Due to the scale's multifaceted character, it provides detailed information on various chatbot's aspects, which is valuable to help chatbot developers improve their chatbots in a targeted manner (Balaji & Borsci, 2019). Balaji and Borsci (2019) provided preliminary evidence for the USIC's validity and reliability, however the scale needs repeated validity and reliability assessment towards standardization.

In this study, we evaluated the USIC's validity and reliability to further the standardization process. Also, we reduced the scale's length to make it more feasible to implement. We performed an extended replication of Balaji and Borsci's (2019) usability study. During the study, participants interacted with multiple chatbots and filled out the USIC and UMUX-Lite after each completed chatbot interaction.

Results showed evidence for the USIC's concurrent validity and reliability, measured by the USIC's factor structure, its relation to the UMUX-Lite and its internal consistency.

The findings suggest that the USIC can fulfil the need for a standardized diagnostic scale to measure user satisfaction with information chatbots. The proposed 14-item USIC is especially promising as it is more compact, making it more efficient and more feasible to implement. The USIC enables researchers and chatbot developers to gain more insight into the user's satisfaction with information chatbots, compare studies and results, and it offers the possibility to improve chatbots in a targeted way.

Keywords: Chatbots, user satisfaction, validity, reliability, standardization

Table of content

List of tables

List of figures

Validity and Reliability of User Satisfaction with Information Chatbots scale

Chatbots are software applications that can simulate human conversations using natural language via text-based messages (Radziwill & Benton, 2017). The user gives input using text to which the chatbot responds by answering in a conversational manner or by performing a requested task (Radziwill & Benton, 2017).

Companies and organisations in various sectors are increasingly using chatbots, for example in education, e-commerce (McTear, Callejas & Griol, 2016), automotive, banking, telecom, energy, insurance (Artificial Solutions Inc., 2020), and healthcare (Beaudry, Consigli, Clark, & Robinson, 2019). Chatbots can help users with a variety of tasks, for example but not limited to, supporting patients with their treatment adherence (Beaudry et al., 2019), improving communication between health care professionals and their patients (Abashev, Grigoryev, Grigorian, & Boyko, 2017), assisting customers with their purchases (Capgemini, 2019), helping file insurance claims by collecting and passing on incident data (Plexal, 2018), and answering customer queries and retrieving information (Jenkins, Churchill, Cox & Smith, 2007).

The chatbot market is predicted to climb from $2.6 billion in 2019 to $9.4 billion by 2024 (Research and Markets, 2019). The rise is not surprising, as implementing chatbots can significantly reduce an organisation’s costs (Capgemini, 2019). For example, Juniper Research (2019) estimated that in 2023 chatbots will save $7.3 billion in operational costs in banking globally, compared to an estimated $209 million in 2019. A survey by Capgemini (2019) also indicated that chatbots are important for the majority of businesses (69%) as they led to a significant cost reduction for customer service (at least 20%) as well as to improved net promotor scores for all companies (i.e., how likely customers would recommend the company based on their experience with the company).

Developments

Chatbots have been around since the 1960s but are getting more attention since 2016 due to the advances in the development of artificial intelligence (AI) (Følstad & Brandtzaeg, 2017). The advances in AI development led to improvements in machine learning and in natural language processing. This resulted in chatbots' capability to communicate with users in a conversational manner in text (Skjuve & Brandtzaeg, 2018), which early chatbots were not yet able to do (Gnewuch, Morana & Maedche, 2017; McTear et al., 2016; Radziwill & Benton, 2017).

At the same time, an increasing number of people started using instant messaging applications in recent years (Gnewuch et al., 2018; McTear, Callejas & Griol, 2016), and became familiar with communicating with the short messages involved in instant messaging. More than 1.5 billion people worldwide used messaging applications in 2017, and in 2019 that number increased to 2.5 billion people (Clement, 2020). Consequently, many potential chatbot users are now used to interacting via instant messaging, likely making it easier for users to learn how to converse with chatbots. The combination of the increasing use of instant messaging and advancements in chatbot technology, led to the increasing interest from companies to deploy chatbots (Gnewuch et al., 2017).

Customer service domain

The interest for chatbots is particularly strong in the customer service domain (Gnewuch et al., 2017). Companies utilize chatbots to function as an automated part of customer service, mainly as an in-between representative that answers questions customers have, as well as helping customers find information on the company's website (Jenkins et al., 2007). Paikari and van der Hoek (2018) define this type of chatbot that retrieves relevant information for its users as information chatbots.

The anticipated benefits of the use of chatbots in customer service are numerous, and apply to both companies and their customers. Customers can receive assistance at any possible instant as chatbots are not restricted to working hours, and customer waiting times can be nearly completely eliminated as chatbots reply instantaneously to customers (Capgemini, 2019; Somasundaram, Kant, Rawat, & Maheshwari, 2019). A benefit for companies is, for example, the chatbot's ability to provide service to many customers simultaneously, without being limited to their employees' working hours. Consequently, a company needs less employees to assist customers, thereby allowing the company to save resources and money (Gnewuch et al., 2017).

User satisfaction

Although chatbots are potentially very beneficial, the anticipated benefits will only be realized if potential users are satisfied with its use and are willing to (continue to) use it. Put differently, users should both accept service by a chatbot and be willing to adopt it (McTear et al., 2016). Various chatbot-driven services have been discontinued due to disappointing usage (Brandtzaeg & Følstad, 2018; Gnewuch et al., 2017), suggesting that users were not satisfied with its use. Additionally, an unsatisfactory chatbot may also cause frustration with its users and may damage the company's image (Brandtzaeg & Følstad, 2018). As such, chatbots need to be continuously improved in order to achieve satisfaction and accomplish continued usage of the chatbots.

To turn disappointing usage around and develop successful chatbots, developers need insight into which chatbot aspects users are satisfied with and which aspects need further improvement. As such, there is a need for a method to properly measure and assess the users' satisfaction levels of their interaction with the chatbot.

Assessing users' satisfaction is a method used for gathering information on the users' experience with systems and products. ISO 9241-11's (2018) description of users' satisfaction includes "the extent to which the user experience that results from actual use meets the user's needs and expectations." ISO 9241-11 (2018) further defines user experience as "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service." Developers can use information on user satisfaction to their advantage in order to improve their chatbot's design. Especially information pertaining to those aspects where modifications have the biggest impact on the user experience is beneficial to -potentially- save time and resources. To gain information on users' satisfaction, developers and researchers need a standardized scale to assess the user satisfaction.

Standardization of scales

As of yet, there are no standardized scales available to assess user satisfaction with chatbots (Balaji & Borsci, 2019). Some researchers attempted to capture user satisfaction with chatbots but did so using non-standardized scales, created to meet the needs for their specific evaluation process (Balaji & Borsci, 2019; Federici et al., 2020). The inconsistent way of testing makes it difficult to evaluate the results and compare between studies and chatbots.

Standardization of scales provide various benefits for companies and researchers. For instance, standardized questionnaires save companies and researchers time, as they do not need to develop a new scale themselves (Berkman & Karahoca, 2016). Rather, they can simply use the already developed standardized questionnaire. Furthermore, standardized questionnaires are easier to replicate. For example, standardized usability questionnaires are found to be more reliable than non-standardized usability questionnaires (Sauro & Lewis, 2016). Also, standardized questionnaires are helpful in collating a series of findings that help

them draw more generalized conclusions, and allow developers or researchers to communicate results more effectively (Berkman & Karahoca, 2016).

Towards standardization, a scale's validity and reliability should be repeatedly confirmed to make sure the scale measures what it claims to measure and the scale's findings are consistent (Kyriazos & Stalikas, 2018). Construct validity is the overarching type of validity (Drost, 2011; Kyriazos, 2018). Construct validity relates to the extent to which variables (e.g., questionnaire items) describe the theoretical latent construct (i.e., factor) that they are developed to measure (Hair, Black, Babin & Anderson, 2010). This includes the internal structure of the scale (Kyriazos, 2018). However, the relation between the scale and the factors cannot be measured directly, due to factors' abstract and latent nature. As such, the relation needs to be evaluated indirectly by measuring the relation between the scale and factor's observable indicators (i.e., questionnaire items). Factor analysis is a method to determine which indicators measure the same factor or factors and form a scale together (Berkman & Karahoca, 2016).

Construct validity requires an accumulation of evidence to substantiate it, such as evidence for criterion validity (Drost, 2011). Criterion validity relates to the extent to which a questionnaire corresponds with one or more external criteria (Drost, 2011). It describes to which extent the questionnaire is in line with different scales that measure similar constructs (Berkman & Karahoca, 2016). One way of evaluating criterion validity is by assessing the scale's concurrent validity; how a questionnaire relates to a priorly standardized scale that is simultaneously conducted (Berkman & Karahoca, 2016; Taherdoost, 2016). The relation between the scale's results indicate to what extent the new questionnaire measures the same (or different) factors.

Reliability relates to how consistent and stable the questionnaire's measurements are (Taherdoost, 2016). One method for evaluating reliability is to assess the questionnaire's

internal consistency (Berkman & Karahoca, 2016). Internal consistency describes the extent to which the questionnaire item's consistency measure the same phenomena and is typically evaluated by using Cronbach's alpha (Drost, 2011).

Another method for showing reliability and stability is by confirming the questionnaire's factor structure in replication (Drost, 2011). Replicating the factor structure, in a different participant population, is a preferred method for showing generalizability (DeVellis, 2016). The factor structure indicates what observations (i.e., questionnaire items) tend to measure the same construct. In subsequent studies it should be evaluated to what extent the measurements of the construct are consistent with the previously found factor structure (Berkman & Karahoca, 2016).

Existing user satisfaction scales

Although there are currently multiple standardized scales available to measure user satisfaction, such as the System Usability Scale (SUS) (Brooke, 1996), the Usability Metric for User Experience (UMUX) (Finstad, 2010) and the UMUX-Lite (Lewis, Utesch & Maher, 2013), these instruments do not focus specifically on chatbots, and fail to reflect all aspects relevant for information chatbots (Tariverdiyeva & Borsci, 2019). Følstad and Brandtzaeg (2017) argue that the design of chatbots differs substantially from, for example, stationary websites. Unlike websites, most of the chatbot's content and features are hidden from the user, and the final design depends on the user's input that contains numerous variations. It is therefore likely that the factors that influence the users' satisfaction are different for chatbots.

Also, the SUS, UMUX, and UMUX-Lite are non-diagnostic in nature (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019). That is to say that these scales show if users are generally satisfied or not, but the scales do not provide information on specific aspects of the user satisfaction and therefore do not reveal what aspects of the system the user is

(un)satisfied with (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019). Without such specific information, developers can only guess how they should improve their product or system. As such, there is a need for a validated diagnostic scale that addresses relevant aspects for chatbots, which is currently not present in existing standardized scales (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019).

Scale for user satisfaction with information chatbots (USIC)

In an effort to create a diagnostic scale specifically for information chatbots, Balaji and Borsci (2019) developed the user satisfaction with information chatbots (USIC) questionnaire. The USIC is a multifaceted scale that indicates the user's satisfaction for different aspects of the chatbot, and which exposes a chatbot's weaknesses and shows its strong suits.

Balaji and Borsci (2019) based their work on the 27 features for the perceived usability of chatbots as identified by Tariverdiyeva and Borsci (2019). Balaji and Borsci (2019) did an initial review of the features' quality and relevance for measuring user satisfaction with information chatbots and they excluded features deemed irrelevant by a focus group. They conducted a literature research and identified three additional relevant features. They then arrived at a list composed of 21 features which are deemed relevant for evaluating the user's satisfaction with information chatbots. They developed three questionnaire items for each of these features, creating a questionnaire consisting of 63 questions. A focus group was used in order to receive feedback on the draft questionnaire, as well as to assess its content adequacy. Participants indicated how relevant they perceived each item to be. Balaji and Borsci (2019) subsequently excluded the irrelevant features and associated items, and finally arrived at a USIC composed of 42 questionnaire items (see Appendix A).

Balaji and Borsci (2019) also conducted a usability study using a group of 60 students to evaluate the 42-item USIC's validity and reliability. They assessed the underlying factor structure and identified a four-factor structure. Waldera and Borsci (2019) used the study's data and identified a nine-factor structure. The first four factors in both structures showed a highly comparable item distribution. However, Waldera and Borsci's (2019) structure excluded two features from the scale and separated five other features into five separate factors, while Balaji and Borsci's (2019) structure included these seven features mainly in the second factor. Balaji and Borsci (2019) based their choice for the four-factor structure on a combination of multiple statistical criteria, meaningful fit of the data and its consistency with their focus group results. Waldera and Borsci (2019) did not provide a rational for their chosen structure. By conducting this study, the researchers made the first step towards standardization. However, the USIC questionnaire needs further psychometric evaluation if this is to be used as a standardized scale.

Effect of age

Research by Moore (2012) shows that individuals from the Millennial and Baby Boom generations have vastly different levels of interactive media usage, such as instant messaging which is involved with chatbot usage. Millennials (i.e., individuals who were born between 1980 and 1995) use interactive media to a significant higher degree and technology is more integrated into their daily lives compared to older individuals (Moore, 2012). Moore (2012) therefore expects that Millennials are better adept to using interactive technology. Based on this, we expect that individuals who are currently between the ages of 25 and 35 are also more adaptive to using chatbots than individuals between 55 and 70 years of age, which likely results in a different experience interacting with the chatbots.

The age groups' different interaction experience, in turn, might affect the USIC's factor structure. For instance, the individuals' communication style could influence Balaji and Borsci's (2019) Communication quality factor that describes "the ease with which the user can initiate an interaction with the chatbot and communicate one's request" (p. 63). Millennials might communicate in a manner that was effective for them during previous interactive technology usage. This type of input might be easier for chatbots to understand than input from older individuals. Older individuals would then likely need to provide more input (e.g., rephrasing, answering clarifying questions), and base their response to the related USIC questions on more input than their initial request only. Consequently, the feature associated with the chatbot's understanding of user input (i.e., Communication effort, see Appendix A, Table A1) may not group with questions related to the conversation's start, such as in the Communication quality factor, and alter the factor structure.

Present study

In this study, we evaluated the USIC's concurrent validity and reliability by performing an extended replication of Balaji and Borsci's (2019) usability study. Similar to the previous study, we conducted a usability study with chatbots and we asked participants to fill out the USIC after their interaction with chatbots. This study differs from Balaji and Borsci's (2019) study, as we included six Dutch chatbots and translated the USIC into Dutch. To gather evidence for concurrent validity, we also included the standardized UMUX-Lite by Lewis, Utesch and Maher (2013) to assess if the USIC measures the same (or different) factors.

The UMUX-Lite is a two-item questionnaire that assesses general user satisfaction in systems. Its brief format is a minimal addition to the session length and helps minimizing the strain on the participants. A moderate to strong correlation between the USIC and UMUX-Lite indicates that the USIC captures the UMUX-Lite's concept.

Moreover, we also explored potential differences in the USIC's factor structure between individuals from two new categories: individuals between 25 and 35 years old and between 55 and 70 years old. So far, Balaji and Borsci (2019) did not take age-related differences into account; they evaluated the USIC with individuals with an average age of 23.7 years (SD = 4.8). Here, we evaluated the USIC's factor structure robustness under the two different age groups.

Furthermore, we assessed if we could create a shortened version of the USIC that addresses all features using a minimal number of questions, whilst maintaining the questionnaire's validity and reliability. Currently the USIC consists of 42 questions, which includes multiple questions per feature to evaluate the user's satisfaction with information chatbots. A shorter and more compact scale that is equally effective, would put less strain on its users by reducing the required time and effort to fill it out (Singh, 2004). As a result, it could potentially increase a user's willingness to fill it out.

The main research questions of this study are related to the validity and reliability of the USIC, and are thus as follows:

RQ1: Is the USIC's factor structure, as identified by Balaji and Borsci (2019), replicable and reliable?

RQ2: Does the USIC show moderate to strong correlations with the UMUX-Lite indicating concurrent validity?

Moreover, associated to our extension of the previous work, we also investigated the following aspects:

RQ3: Does the factor structure differ substantially for individuals between 25 and 35 years old compared to individuals between 55 and 70 years old?

RQ4: Can we create a shortened version of the USIC that addresses all relevant features as identified by Balaji and Borsci (2019)?

Method

USIC and UMUX-Lite translation

Before conducting the test, we translated the USIC questionnaire and UMUX-Lite into Dutch to optimize the participants' comprehension. To ensure the quality of the translation, the Dutch version of the questionnaires was translated back into English by two individuals who are fluent in both English and Dutch. We compared both translations with the original version, and any identified differences were highlighted and discussed with the translator concerned. After this consultation round, we made a total of 11 changes (see Appendix A, Table A2). Notably, both translators were unaware that another translator translated the questionnaires also, as to not influence their work.

Participants

A total of 60 participants participated in the study. The population consisted of 30 individuals between 25 and 35 years old ($M = 28.80$, $SD = 2.70$), and 30 individuals between 55 and 70 years old ($M = 62.30$, $SD = 3.89$).

All participants indicated that they had at least a basic understanding of English in terms of reading and writing; one participant had a basic understanding of English, twelve participants had a moderate understanding, forty participants had a good understanding of the language, and seven possessed an excellent understanding of English.

Recruitment

We recruited participant based on the following four criteria:

- The individuals had to be between 25 and 35 or 55 and 70 years of age.
- The individuals needed to have a good understanding of the Dutch language.

- The individuals needed to have at least a basic understanding of the English language, in terms of reading and writing.
- The individuals had to have access to a computer with internet capabilities in order to participate in the study.

Participants were recruited using the snowball technique. We reached out to potential participants using some basic information on the study's goals, activities, duration, and method of conducting. If individuals indicated they were interested in participating, we provided them with more detailed information and subsequently scheduled an appointment. After scheduling this, we sent the participant an e-mail with the scheduled time and date, the information sheet, the informed consent form and information on the video-connection platform that was to be used.

Procedure

Due to the limitations imposed by the COVID-19 pandemic, the test sessions had to be conducted online using a video connection. The participants were asked to share their computer screen when starting with the chatbot tasks. The session administrator used a webcam to make participants feel at ease and assisted with any non-task-related technical difficulties.

Each participant joined an online session of one to one and a half hours. The session administrator welcomed the participant via a video connection and briefly explained the study's goal and the session activities. The session administrator then explained to the participants that they would have to do a task with a chatbot, after which they would receive a questionnaire asking for their feedback on their experiences with the chatbot (see Appendix B for the session script).

The session administrator asked the participant to read and sign the informed consent form on Qualtrics prior to starting the activities (see Appendix B for the informed consent form). The informed consent form explained the study's goal, the session activities, what data would be collected, confidentiality and potential risks. Also, the informed consent form asked the participants' permission for audio and screen recording, and reiterated that the participant could stop the session at any time. The form mentioned the university's ethical approval, and listed the researcher's contact information. Participants could only participate in the study after agreeing to all consent questions.

The session administrator subsequently asked the participant to fill out a short demographic questionnaire on the participant's age, their Dutch and English language proficiency, their highest completed level of education, and their previous experiences with chatbots.

The session administrator subsequently oriented the participant to chatbot-related tasks and questionnaires. Each participant performed tasks using five chatbots (see Appendix B for all chatbots). For each chatbot, the participant received a use scenario and a task. After completing the task, the participant had to fill-out the USIC and UMUX-Lite for the associated chatbot based on his or her experience. At the end of the session, the session administrator answered any remaining questions the participant had, thanked the participant and ended the session.

We semi-randomly assigned five chatbots to each participant, using Qualtrics survey software randomisation tool. Specifically, we randomly assigned two English chatbots previously tested in Balaji and Borsci (2019) and three Dutch chatbots to each participant. We counterbalanced the assignments to achieve an equal distribution and enhance the study's internal validity. Additionally, we randomized the questionnaire item sequence.

The session administrator directed the participant to the chatbot if it took a participant more than one minute to locate the chatbot on the website. This situation occurred several times, in particular with the KPN and Absolut chatbots. The session administrator noted each assistance occurrence in the session notes.

If, after interacting with the chatbot, a participant considered a task impossible to complete, he or she could continue to fill out the USIC questionnaire. The session administrator noted these cases in the session notes.

Materials

We used the following materials for each session: a computer with an internet connection, microphone, Flashback Express Player for audio- and screen-recording, Qualtrics to present participants with the informed consent form, chatbot tasks, translated USIC, translated UMUX-Lite, a video connection using Whereby, Microsoft Excel for note taking, a session administrator script, an informed consent form and a document explaining participants how to set up the video connection.

We included a set of ten chatbots in the study: four English chatbots, previously included in Balaji and Borsci's study (2019) (e.g., Australian Taxation Office) and six new Dutch chatbots (e.g., Bol.com). The complete list of chatbots and the associated URLs can be found in Appendix B, Table B2. Notably, rather than directing the participants to chatbot's specific webpage, we provided participants with the general website URL and had them look for the chatbot.

After the participants completed the demographic questionnaire, we asked them to complete an information retrieval task, similar to the tasks included in the Balaji and Borsci (2019) study. Participants received a short use scenario and task for each chatbot they interacted with. We designed the chatbot task to be representative for use on that particular

website. For example, we included the following task in Dutch for a chatbot of an energy and gas supplier: *"You're considering switching to Oxxio's green energy. However, the contract with your current energy supplier has not yet ended, and your energy supplier will impose a cancellation penalty if you switch suppliers before the end date. You want to use the chatbot to find out whether Oxxio will pay this fine for you if you switch to Oxxio"* (see Appendix B for all chatbot tasks).

In case of an English chatbot, the participants received the task both in Dutch and in English to help participants formulate their request. See Appendix B, Table B3 for the task prompts for all chatbots.

To gather evidence for concurrent validity, we included the standardized UMUX-Lite by Lewis, Utesch and Maher (2013) for user satisfaction to compare the USIC's results with. The UMUX-Lite is a two-item questionnaire that assesses general user satisfaction in systems. Its brief format was a minimal addition to the session length and helped minimizing the strain on the participants.

Results

Data set preparation

The dataset consisted of one data line per chatbot and participant combination. Each of the 60 participants interacted with five chatbots. Four incomplete data lines were removed due to incomplete answers, resulting in a dataset containing 296 lines of data. The negatively worded questionnaire item scores (i.e., Q10 and Q11) were inverted before performing the analysis.

USIC's factor structure

To assess the USIC's factor structure, a principal component analysis (PCA) was conducted on the questionnaire's 42 items. First, all three PCA assumptions were assessed to

establish if the use of the PCA was appropriate for the current dataset. The correlation matrix showed that all items had at least one correlation greater than 0.3. The Kaiser-Meyer-Olkin (KMO) measure for sampling adequacy showed an overall value of 0.927, and the values of all individual items were greater than 0.7, indicating a more than acceptable adequacy according to Kaiser (1974). The Bartlett's Test of Sphericity was statically significant ($p < .001$), which indicated sufficiently large relations between items in order to be able to conduct the PCA (Field, 2009). As such, all assumptions for the PCA were met and it was acceptable to continue.

Subsequently, the PCA was conducted. Usually researchers use a criterion as input for a first attempt to interpret a certain factor structure, and assess whether the factor structure can be interpreted meaningfully (Hair et al., 2010). One of such considerations, is the number of factors based on prior research. Here, Kaiser's criterion of one and the scree plot were used as criteria for initial assessment and interpretation.

The PCA results showed eight factors with eigenvalues greater than Kaiser's criterion of one. Visual inspection of the scree plot showed an inflection point at two factors (see for Appendix C, Figure C1 for the scree plot).Together, these results suggested that the number of factors to be retained, is most likely to be between two and eight, which approaches the factor range of three to seven factors identified by Balaji and Borsci (2019). After further analysis they arrived at their four-factor structure. Noting that the factor range found in this study neared the range found by Balaji and Borsci (2019) and, based on their work, we continued to evaluate the four-factor structure.

To further assess the four-factor structure, additional PCA's factor indicators were addressed. The four factors explained 57.6% of the total variance and 35.6%, 10.9%, 6.2%, 4.8% of the individual variances. A total explained variance of 50 to 60% is considered satisfactory in social sciences (Hair et al., 2010; Pett, Lackey, Sullivan, 2003). As such, the

four-factor structure's total variance was adequate. The Varimax orthogonal rotation was conducted for the interpretation of the factors and indicated a simple structure. That is, the items loaded strongly onto only one factor, suggesting an optimal structure (see Appendix C, Table C4 for the factor loadings of the 42-item USIC) (Hair et al., 2010; Thurstone, 1947). The factors showed a meaningful item distribution that showed great consistency with the distribution as also identified by Balaji and Borsci (2019) (see Table 1).

Table 1.
The factor structure of the 42-item USIC identified by Balaji and Borsci (2019) and the present study, showing the items included in each factor and the item's associated features.

F#	Factor structure 42-item USIC Balaji and Borsci (2019)		Factor structure 42-item USIC present study		Associated feature
	Factor name	Items	Factor name	Items	
F1	Communication quality	Q1, Q2, Q3,	Conversation start	Q1, Q2, Q3,	Ease of Starting a Conversation,
		Q4, Q5, Q6,		Q4, Q5, Q6	Accessibility,
		Q10, Q11		n/a	Communication Effort
F2	Response quality	Q7, Q8, Q9,	Communication	Q7, Q8, Q9,	Expectation setting,
		Q12,	quality	**Q10**, **Q11***, Q12,	Communication effort,
		Q14, Q15,		**Q13**, Q14, Q15,	Maintain themed discussion,
		Q16, Q17, Q18,		Q16, Q18,	Reference to service
		Q22, Q23, Q24,		Q22, Q23, Q24,	Recognition and facilitation of user's goal & intent,
		Q25, Q26, Q27,		Q25, Q26, Q27,	Relevance,
		Q28, Q29, Q30,		Q28, Q29, Q30,	Maxim of quantity,
		Q31, Q32, Q33,		Q31, Q33*,	Graceful breakdown,
		Q34, Q35, Q36,		Q34, Q35,	Understandability,
		Q37, Q38, Q39		Q37, Q39	Perceived credibility
F3	Perceived privacy	Q13,	Perceived privacy	n/a	Maintain themed discussion,
		Q19, Q20, Q21		Q19, Q20*, Q21,	Perceived privacy,
		n/a		**Q32*,**	Graceful breakdown,
		n/a		**Q38***	Perceived credibility
F4	Perceived speed	n/a	Perceived speed	**Q36***,	Understandability
		Q40, Q41, Q42		Q40*, Q41, Q42	Perceived speed

Note. The table shows the items of one feature per row. * items removed during item selection to the refined 33-item USIC. Items differences compared to Balaji and Borsci (2019) in boldface

The USIC's internal consistency was evaluated using Cronbach's alpha. The USIC had a very high internal consistency, with a Cronbach's alpha of 0.948. Also, the individual

factors separately had high internal consistency ratings with $\alpha = 0.918$ for factor 1 (F1), $\alpha = 0.961$ for factor 2 (F2), $\alpha = 0.731$ for factor 3 (F3), and $\alpha = 0.767$ for factor 4 (F4). The very high internal consistency therefore allowed for item reduction and optimisation of the USIC as envisioned in our second objective.

Item selection

One of the study's aims was to create a shortened version of the USIC that addresses all features using a minimal number of questions, whilst maintaining the questionnaire's validity and reliability. First, the USIC was refined by iteratively evaluating and omitting items based on its factor loading, Cronbach's alpha if an item was deleted, and corrected item-total correlations, respectively. Items with a factor loading greater than 0.5 were considered practically significant and were retained (Hair et al., 2010). To further optimize the questionnaire's internal consistency, and thus reliability, items that lead to an increase in Cronbach's alpha when deleted, or items with a corrected item-total correlation below 0.5 were removed (Hair et al., 2010). Cronbach's alpha if an item was deleted and the corrected item-total correlations were computed per factor. A total of nine items were removed from the dataset following this procedure. Five items (Q9, Q17, Q32, Q33, Q38) had a factor loading less than 0.5, three items (Q20, Q36, Q40) showed an increase of Cronbach's alpha if deleted, and one item (Q11) showed a corrected item-total correlation below 0.5 in combination with a slightly increased Cronbach's alpha. Removal of these items resulted in a 33-item list and in the refinement of factors 2, 3, and 4. The 33-item USIC had a very high internal consistency with $\alpha = 0.946$ for the entire questionnaire, with F1 $\alpha = 0.918$, F2 $\alpha = 0.962$, F3 $\alpha = 0.879$, and F4 $\alpha = 0.916$.

Although these 33 questions provide for a good questionnaire, there is still the possibility for further refinement. The 33-item list included multiple items per feature (see

Table 1). Asking users to fill out only one question per feature would reduce the questionnaire's length substantially (i.e., from 33 to 14 items), which would be more efficient and put less strain on users, potentially increasing user's willingness to fill it out. As such, it was decided to further reduce the number of items and retain those items with the highest factor loading for each feature as those items show the strongest relationship with the underlying latent factor and preserve the factor's reliability (Bollen & Lennox, 1991).

As a result, 14 items were retained (see Table 2), making the USIC more efficient to fill out and thus more feasible to implement. Concurrent validity was indicated by the internal correlations. The majority of factor 1 and 2's internal correlations were greater than 0.5, and all were at least greater than 0.3 except for one correlation; the correlation between Q10 and Q37 was 0.271. Factors amongst each other showed weak correlations ($r > .3$) (see Appendix C, Table C3 for the correlation matrix of the optimized 14-item USIC).

Cronbach's alpha for the refined 14-item USIC questionnaire was $\alpha = 0.874$, indicating a high reliability. Cronbach's alpha for factors 1 and 2 separately were $\alpha = 0.778$ and $\alpha = 0.919$, respectively. Factors 3 and 4 only contained a single item so Cronbach's alpha could not be calculated.

Although single-item factors are generally discouraged, there are exceptions. Factors may have a simple and narrow definition that can be adequately covered by a single item (Hair et al., 2010). A single item can suffice if the meaning is clear, easily understandable and distinct. It was argued that a single item was sufficient for factors 3 (Q19 and Q21) and factor 4 (Q41 and Q42) as the items for both factors ask direct questions about the factor's content and the items have a high resemblance in meaning.

Table 2.
The 14-item USIC composed of the items with the highest factor loading for each feature, and each item's associated feature and factor loadings.

Q#	Question	Feature	F1 Conversation start	F2 Communication quality	F3 Perceived Privacy	F4 Perceived speed
Q2	It was easy for me to understand how to start the interaction with the chatbot.	Ease of starting a conversation	0.820	0.059	0.006	0.163
Q5	The chatbot function was easily detectable.	Accessibility	0.904	0.001	0.057	-0.067
Q7	Communicating with the chatbot was clear.	Expectation setting	0.234	0.709	0.093	0.122
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me. (R)	Communication effort	0.002	0.627	-0.022	-0.213
Q15	The chatbot maintained relevant conversation.	Ability to maintain themed discussion	0.067	0.858	0.057	0.106
Q16	The chatbot guided me to the relevant service.	Reference to service	0.065	0.763	-0.052	0.133
Q19	The interaction with the chatbot felt secure in terms of privacy.	Perceived privacy	0.124	0.138	0.906	0.112
Q24	I find that the chatbot understands what I want and helps me achieve my goal.	Recognition and facilitation of user's goal and intent	0.006	0.878	0.113	0.031
Q27	The chatbot provided relevant information as and when I needed it.	Relevance	0.076	0.874	0.030	0.096
Q29	The chatbot gives me the appropriate amount of information.	Maxim of quantity	-0.065	0.785	-0.013	0.182
Q31	The chatbot could handle situations in which the line of conversation was not clear.	Graceful breakdown	-0.015	0.704	0.079	0.085
Q34	I found the chatbot's responses clear.	Understandability	0.109	0.664	0.131	0.285
Q37	I feel like the chatbot's responses were accurate.	Perceived credibility	0.103	0.625	0.151	0.322
Q42	The chatbot is quick to respond.	Perceived speed	0.084	0.130	0.044	0.876

Comparative analysis

To assess the factor structure in more detail, this study's item distribution was compared with the item distribution found by Balaji and Borsci (2019). A total of 35 out of the 42 items were similarly distributed over the four factors compared to Balaji and Borsci's (2019) findings. Six other items out of the 42 items loaded in the current study onto a different factor than in the study of Balaji and Borsci (2019), and the remaining one item (Q17) did not load on any factor (see Table 3). Notably, five of these seven last mentioned

items (Q11, Q17, Q32, Q36, Q38) were removed here during refinement due to low factor loadings. The other two items (Q10, Q13) loaded onto the present study's Communication quality factor (F2), causing these to be grouped with the items of the associated features.

Table 3.
USIC items that loaded on a different factor in the present study when compared with Balaji and Borsci (2019)

		Item's factor location	
Q#	Question	Balaji and Borsci (2019)	Present study
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me.	F1 Communication quality	F2 Communication quality
Q11*	I had to pay special attention regarding my phrasing when communicating with the chatbot.	F1 Communication quality	F2 Communication quality
Q13	The interaction with the chatbot felt like an ongoing conversation.	F3 Perceived privacy	F2 Communication quality
Q17*	The chatbot is using hyperlinks to guide me to my goal.	F2 Response quality	None
Q32*	The chatbot explained gracefully when it could not help me.	F2 Response quality	F3 Perceived privacy
Q36*	The chatbot's responses were easy to understand.	F2 Response quality	F4 Perceived speed
Q38*	I believe that the chatbot only states reliable information.	F2 Response quality	F3 Perceived privacy

Note. * Items that were removed during refinement process towards 33-item USIC due to a factor loading below 0.5

Correlation UMUX-Lite and USIC

To assess the USIC's concurrent validity, the correlation between the USIC and UMUX-Lite was examined. For each data line mean scores were calculated for the UMUX-Lite and USIC. The correlations between the 33-item and 14-item USIC and UMUX-Lite were estimated using Spearman's rank-order correlation. Both USIC versions showed a strong correlation with the UMUX-Lite, as can be seen in Table 4, indicating concurrent validity for the overall questionnaire.

When looking at the factors separately, it could be seen that factor 2 of both questionnaires also showed a strong correlation. That said, factors 1 and 4 of both USICs showed very weak correlations with UMUX-Lite. Factor 3 of the 33-item USIC showed a

weak correlation and the correlation between the 14-item USIC's and the UMUX-Lite was not significant.

Table 4.
Correlations between UMUX-Lite and the 33-item and 14-item USIC

	UMUX-Lite
33-item USIC	.837*
(F1) Conversation start factor	.288*
(F2) Communication quality factor	.804*
(F3) Perceived privacy factor	.306*
(F4) Perceived speed factor	.259*
14-item USIC	.821*
(F1) Conversation start factor	.266*
(F2) Communication quality factor	.794*
(F3) Perceived privacy factor	.286 Ns
(F4) Perceived speed factor	.223*

Note. Ns = not significant, *ρ<.001

Differences for the two age categories

The USIC's factor structure of the individuals between 25 and 35 years of age (25-35 group) and individuals between 55 and 70 years of age (55-70 group) was compared to see whether a substantial difference existed (see Table 5). An identical procedure to the assessment of the overall USIC's factor structure was followed.

All assumptions for the PCA were met for both age groups after removing Q17 for the 25-35 group. The correlation matrix showed that Q17 correlates lowly with all the other items ($-0.3 < r < 0.3$). After removal of Q17 for the 25-35 group, all items for both age groups showed a correlation greater than 0.3. Both the 25-35 and 55-70 group, had a high overall KMO (0.862 and 0.897, respectively), and the individual KMO was above 0.6. Also, both groups passed the Bartlett's Test of Sphericity ($p < .001$) (Field, 2009). As such, all assumptions for the PCA were met and it was acceptable to continue.

As indicated in Table 5, the PCA results suggested a meaningful fit for the four-factor structure due to the combination of the range indicated between the factors with eigenvalue

greater than one, the scree plot inflection point, the adequate variance explained by four factors (i.e., greater than 50%), the simple structure, and the groups showed a meaningful item distribution as indicated in Table 6.

Table 5.
The PCA results of the four-factor structure and its internal consistency for the 25-35 group and 55-70 group

PCA indicators		25-35 group	55-70 group
Factors with eigenvalues greater than one		8 factors	8 factors
Scree plot inflection point		3 factors	2 factors
Total variance explained by 4 factors		56.2%	61.2%
Individual variance explained per factor		31.7%, 12.1%, 7.4%, and 4.8%	39.5%, 10.3%, 5.9%, and 5.5%
Varimax orthogonal rotation		Simple structure with some weak cross loadings	Simple structure
Cronbach's alpha	Overall	0.934	0.948
	(F1) Conversation start	0.926	0.918
	(F2) Communication quality	0.952	0.962
	(F3) Perceived privacy	0.815	0.801
	(F4) Perceived speed	0.910	0.856

The factors showed a meaningful item distribution which was consistent with the majority of the distribution of the complete dataset (see Table 6). However, for the 25-35 group, the items that belong to the features Understandability (Q34, Q35, Q36) and Perceived credibility (Q37, Q38, Q39) loaded on factor 3 instead of factor 2.

Table 6.
The USIC's item distribution, before refinement, of the current study's complete participant group, 25-35 group, 55-70 group, compared to the item distribution identified by Balaji and Borsci (2019),

		Current study		
	Balaji and Borsci (2019)	Complete participant group	25-35 group	55-70 group
F1	Q1, Q2, Q3, Q4, Q5, Q6, Q10, Q11	Q1, Q2, Q3, Q4, Q5, Q6	Q1, Q2, Q3, Q4, Q5, Q6	Q1, Q2, Q3, Q4, Q5, Q6
F2	Q7, Q8, Q9, Q12, Q14, Q15, Q16, Q17, Q18, Q22, Q23, Q24,	Q7, Q8, Q9*, Q10, Q11**, Q12, Q13, Q14, Q15, Q16, Q18, Q22, Q23, Q24,	Q7, Q8*, Q10, Q11, Q12, Q13*, Q14, Q15, Q16, Q18, Q22, Q23, Q24,	Q7, Q8, Q9*, Q10, Q11*, Q12, Q13, Q14, Q15, Q16, Q18, Q22, Q23, Q24,

		Current study		
	Balaji and Borsci (2019)	Complete participant group	25-35 group	55-70 group
	Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q32, Q33, Q34, Q35, Q36, Q37, Q38, Q39	Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q33*, Q34, Q35, Q37, Q39	Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q33*	Q25, Q26, Q27, Q28, Q29, Q30, Q31, Q32*, Q33, Q34, Q35, Q36, Q37, Q38*, Q39
F3	Q13, Q19, Q20, Q21	Q19, Q20**, Q21, Q32*, Q38*	Q9*, Q19*, Q21, Q34, Q35, Q36, Q37, Q38, Q39	Q19, Q20, Q21
F4	Q40, Q41, Q42	Q36**, Q40**, Q41, Q42	Q40**, Q41, Q42	Q40, Q41, Q42

Note. The table shows the items of one feature per row.

* Items removed during refinement because of factor loading below 0.5

** Items removed during refinement because of improving Cronbach's alpha or corrected item-total correlation

Item selection age categories

The same procedure of items selection as for the total participant group was employed for the age groups. For the 25-35 group, a total of eight items were removed from the dataset. Seven items (Q8, Q9, Q13, Q19, Q20, Q32, Q33) had a factor loading less than 0.5, and two items (Q21, Q40) showed an increase of Cronbach's alpha when deleted. Although Q21 showed an increase in Cronbach's alpha when deleted, it was decided not to remove the item because Q21 was the only remaining representation of the Perceived privacy feature. Removal of the eight items resulted in the refinement of factors 2, 3, and 4.

For the 55-70 group, a total of eight items were removed from the dataset following this procedure. Five items (Q9, Q11, Q17, Q32, Q38) had a factor loading less than 0.5, one item (Q20) showed an increase of Cronbach's alpha when deleted, and two items (Q33, Q40) had a corrected item-total correlation below 0.5. Removal of the eight items resulted in the refinement of factors 2, 3, and 4.

For each feature, the items with the highest factor loading were selected from the refined item list, and this resulted in the questionnaire structures as outlined in Table 7.

Table 7.
The USIC items with the highest factor loading per feature for the complete participant group, the 25-35 group and 55-70 group

	Items with highest factor loading		
Feature	Complete participant group	25-35 group	55-70 group
Ease of starting a conversation	Q2	Q2	**Q1**
Accessibility	Q5	**Q6**	Q5
Expectation setting	Q7	Q7	Q7
Communication effort	Q10	Q10	**Q12**
Ability to maintain themed discussion	Q15	Q15	Q15
Reference to service	Q16	Q16	Q16
Perceived privacy	Q19	**Q21**	Q19
Recognition and facilitation of user's goal and intent	Q24	Q24	Q23
Relevance	Q27	Q27	Q27
Maxim of quantity	Q29	Q29	**Q30**
Graceful breakdown	Q31	Q31	Q31
Understandability	Q34	Q35	**Q34**
Perceived credibility	Q37	**Q39**	Q37
Perceived speed	Q42	**Q41**	Q42

Note. Items that differ from complete participant group are indicated in boldface

For eight features a different item was suggested for one of the two age groups when compared to the total participant group (see Table 7). For six items, the difference in factor loading between an age group and the total participant group was minimal (i.e., below 0.02). The difference in factor loading for the items associated with the features Understandability and Perceived credibility showed a somewhat greater difference, but were still quite small with differences of 0.103 and 0.053, respectively.

All three 14-item USICs showed a high internal consistency under its corresponding population (see Table 8). Cronbach's alpha could not be calculated for factors 3 and 4 because these factors consisted of a single item in all three 14-item USICs.

Table 8.
Cronbach's alpha for the 14-item USICs and its four factors for the complete participant group, 25-35 group, and 55-70 group

	Cronbach's alpha		
Feature	Complete participant group	25-35 group	55-70 group
Complete 14-item USIC	.874	.848	.905
(F1) Conversation start factor	.778	.773	.760
(F2) Communication quality factor	.919	.898	.943
(F3) Perceived privacy factor	n/a	n/a	n/a
(F4) Perceived speed factor	n/a	n/a	n/a

Discussion

The present study conducted a psychometric evaluation of the USIC questionnaire's validity and reliability using a new population of individuals between 25-35 and 55-70 years old. The data showed a meaningful fit for Balaji and Borsci's (2019) four-factor structure and the item distribution showed great similarity with Balaji and Borsci's (2019) findings as well. The complete USIC as well as its four factors had high internal consistency, showing high reliability. The UMUX-Lite strongly correlated with the complete USIC and the present study's Communication quality factor (F2), providing support for concurrent validity.

Factor structure

The first research question was "Is the USIC's factor structure, as identified by Balaji and Borsci (2019), replicable and reliable?" To answer the research question, we performed a PCA. The results showed that the data supports the four-factor structure of Balaji and Borsci (2019), thus providing evidence for a similar internal structure and its structural stability (Kyriazos, 2018). Notably, the four-factor structure explained 57.6% of the total variance. According to Hair et al. (2010) and Pett et al. (2003) 50 to 60% is considered satisfactory in social sciences as information is less precise compared to natural sciences, that use more exact measurements and where an explained total variance level of 95% is considered

appropriate. Although here 57.6% is considered adequate, it should be born in mind that 42.4% of the total variance was not explained by the four-factor structure, which suggests that the questionnaire could be further optimized for more comprehensiveness.

Moreover, the four-factor structure is supported by the meaningful item distribution, which is similar to Balaji and Borsci's (2019) distribution for the majority of the items (see Table 1). Also, by replicating and confirming Balaji and Borsci's (2019) results under a new population, we provided evidence for generalizability (DeVellis, 2016).

Revised item's distribution

The results showed that the items Q10 and Q13 were distributed differently compared to Balaji and Borsci (2019) and were loaded onto the present study's Communication quality factor (F2) instead of Conversation start factor (F1). We argue that these items have a better and more meaningful fit in the present study than in the study by Balaji and Borsci (2019) (see Table 9) for the following reasons:

- **Q10.** Q10 asks about the need for rephrasing, which we argue is more in line with the Communication quality factor's content (F2) than that of the Conversation start factor (F1). The features in the Communication quality factor describe how well a chatbot performs in the communication aspects of the interaction (see Appendix A, Table A). In Balaji and Borsci's (2019) work, the item was grouped with features that highlighted the Conversation's start (i.e., Ease of starting a conversation, and Accessibility, see Table 1). However, rephrasing was not limited to the Conversation start in the present study, but instead this happened throughout the complete interaction.
- **Q13.** Similarly, we argue that Q13 provides a better fit onto the Communication quality factor (F2) instead of Balaji and Borsci's (2019) proposed fit onto the Perceived privacy factor (F3). Q13 asks users the extent to which the interaction felt

like an ongoing conversation (see Appendix A, Table A1). As such, the item's content does not seem to be directly associated with how well users feel their privacy is protected. Instead, this item seems be associated with the quality of the chatbot's response, which is captured in the Communication quality factor (see Table 9).

Factor interpretation

The slight difference in item distribution (see Table 1) led us to reinterpret factors for the refined USIC (see Table 9). Based on this study's data, we reinterpreted factor 1 and 2 as follows: (F1) Conversation start, or the ease with which the user can access the chatbot and start the interaction, and (F2) Communication quality, or the chatbot's ability to understand the user's input and the quality of the chatbot's response to it. The difference in factor interpretation is mainly caused by item Q10. We interpreted factors 3 (Perceived privacy) and 4 (Perceived speed) the same as Balaji and Borsci (2019) did, as these factors had the main focus on the items included in the present study (see Table 1).

Table 9
Factor interpretation of USIC in Balaji and Borsci's (2019, page 63) study and the present study

F#	Balaji and Borsci (2019, page 63)		Present study	
	Factor name	Interpretation	Factor name	Interpretation
F1	Communication quality	"The ease with which the user can initiate an interaction with the chatbot and communicate one's request"	Conversation start	The ease with which the user can access the chatbot and start the interaction.
F2	Response quality	"The quality of the response provided by the chatbot after the user has provided some form of input"	Communication quality	The chatbot's ability to understand the user's input and the quality of the chatbot's response to it.
F3	Perceived Privacy	"The extent to which the user feels that their privacy is being protected during the interaction"	Perceived Privacy	The extent to which the user feels that their privacy is being protected during the interaction"
F4	Perceived Speed	"How quickly the chatbot seems to respond to a given input"	Perceived Speed	How quickly the chatbot seems to respond to a given input

Reliability assessment by internal consistency

In our first research question, we also asked whether the factor structure was reliable. The results showed that Cronbach's alpha was high to very high for the overall questionnaire.

This also applied to each of the USIC's factors in both the unrefined 42-item and in the refined 33-item versions (Field, 2009). As such, the current study's USIC, and its factors, showed good internal consistency, which indicates that the USIC used is a reliable scale.

Concurrent validity UMUX-Lite and USIC

Our second research question was *"Does the UMUX-Lite show a moderate to strong correlation with the USIC?"* The results showed that the UMUX-Lite had a strong relation with both the 33-item and 14-item USIC and the USIC's Communication quality factor (F2). The relations indicate that UMUX-Lite's concept of user satisfaction is captured within the questionnaire and, more specifically, within the USIC's Communication quality factor (F2). The UMUX-Lite's weak to very weak correlation with the factors Conversation start (F1), Perceived privacy (F3), and Perceived speed (F4) suggest that these factors measure a different aspect of the user satisfaction.

That the UMUX-Lite was not reflected in all USIC's relevant factors is directly in line with previous findings by Tariverdiyeva and Borsci (2019) and Waldera and Borsci (2019). Tariverdiyeva and Borsci (2019) found that the UMUX-Lite only measured their Perceived ease of use feature. In Waldera and Borsci's (2019) study, the UMUX-Lite strongly related to their 25-item USIC and to some, but not all, of the features. They identified a strong relation between the UMUX-Lite and the features Reference to service, Recognition of user's intent and goal, Perceived credibility, and the Ability to maintain themed discussion, which are all included in this study's Communication quality factor (F2). Other features showed only a weak or moderate relation with the UMUX-Lite in Waldera and Borsci's (2019) study. The consistent findings imply that the UMUX-Lite's overall user satisfaction concept is reflected within a segment of the USIC.

We argue that the USIC's diagnostic character is a logical explanation for the UMUX-Lite's weak relation with the factors Conversation start (F1), Perceived privacy (F3), and Perceived Speed (F4). The UMUX-Lite is a general assessment of user satisfaction with systems (Lewis et al., 2013). The USIC is designed to provide a more complete picture of the user's satisfaction and assesses additional aspects of the interaction (Balaji & Borsci, 2019). Also, considering the USIC's foundation in literature, and its evaluation by an expert panel and focus group (Balaji & Borsci, 2019; Tariverdiyeva & Borsci, 2019), we consider it reasonable to assume that the USIC provides a more elaborate evaluation, and that its factors Conversation start, Perceived privacy, and Perceived speed are valuable additional features that supports the USIC's diagnostic character and should therefore be retained.

Age groups

We asked in the third research question whether the factor structure for the two separate age categories (i.e., individuals between 25 and 35 years old and between 55 and 70 years old) differed substantially. The results showed a four-factor structure for both groups and the item distribution also showed a great similarity except for the items related to two features. The items associated with the features two Understandability and Perceived credibility (i.e., Q34, Q35, Q36, Q37, Q38, Q39, see Table 10) loaded for the younger participants onto the Perceived privacy factor (F3), while for the older participants, as well as for the complete participant group, these features were loaded onto the Communication quality factor (F2).

Table 10.
USIC items that loaded onto the Perceived privacy factor (F3) for the 25-35 group

Q#	Question	Associated feature
Q34	I found the chatbot's responses clear.	Understandability
Q35	The chatbot only states understandable answers.	
Q36	The chatbot's responses were easy to understand.	
Q37	I feel like the chatbot's responses were accurate.	Perceived credibility
Q38	I believe that the chatbot only states reliable information.	
Q39	It appeared that the chatbot provided accurate and reliable information.	

The difference suggests that the participants between 25 and 35 years of age have a different association with the features Understandability and Perceived credibility than the older participant group, and have an underlying latent factor that is different from the Perceived privacy factor (F3). As such, we reinterpreted the factor that is composed of the features Understandability, Perceived credibility and Perceived privacy as being Trustworthiness, or the extent to which the user is able to trust the chatbot to provide accurate and understandable information.

Previous analyses by Balaji and Borsci (2019) and Waldera and Borsci (2019) did not identify a factor similar to this study's Trustworthiness factor. Remarkably, the participants in their studies had an average age of 23.7 years (SD= 4.8) and were thus close to the age of the younger participant group in this study. This implies that the participant's age is not the constant factor, and indicates that it may not have been the explanatory factor here. When looking at the participant demographics (see Appendix C, Table C1), it is notable that most younger participants had used a chatbot before (n=29), whereas only half of the older participants reported having used a chatbot prior to this study (n=16). It may be that these younger participants' prior usage affected their interactions with chatbots, resulting in a different factor structure. In earlier research, Borsci, Federici, Bacci, Gnladi and Bartolucci (2015) found that the dimensionality of the SUS and UMUX-Lite was affected by the level of prior experience. Considering the USIC's relation with the UMUX-Lite, the findings may

indicate that the USIC measures different underlying factors for users with prior experience using chatbots. That said, these younger participants stated that they did not have much experience, as 26 out of the 29 participants stated that they only rarely used chatbots.

Optimized 14-item USIC

Our fourth research question was "*Can we create a shortened version of the USIC that addresses all relevant features as identified by Balaji and Borsci (2019)?*" We retained the items with the highest factor loading for each feature to address all features using a minimal number of questions, and arrived at the 14-item USIC as described in Table 11.

Evidence for the 14-item USIC's validity and reliability was provided by its similar results to the refined USIC. Specifically, the 14-item USIC's similar factor structure and item distribution, as compared to the 33-item USIC, indicates that the factor's meaning did not change after removing the items. The strong relation between the overall USIC and the Communication quality factor indicates that UMUX-Lite's concept of user satisfaction is captured within the questionnaire. The high Cronbach's alpha showed internal consistency and, thus, reliability for the overall 14-item USIC, and its Conversation start and Communication quality factors.

The optimized USIC thus enhances the questionnaire's efficiency as it avoids repetition (i.e., it does not address features multiple times) while it is still equally effective by addressing all relevant aspects for user satisfaction with information chatbots. The reduced scale requires less effort and time for users to fill out due to its compact size and thus reduces the strain on its users (Singh, 2004). Several of the participants commented that they felt the scale repeated questions, and a few participants wondered out loud if it was necessary to have highly similar questions included in the questionnaire. This indicates potential users favour a

shorter questionnaire and the 14-item USIC's shorter length could potentially increase users' willingness to fill it out.

Table 11.
The optimized 14-item USIC and each question's associated factor and feature

F#	Factor name	Feature	Q#	Question
F1	Conversation start	Ease of starting a conversation	Q2	It was easy for me to understand how to start the interaction with the chatbot.
		Accessibility	Q5	The chatbot function was easily detectable.
F2	Communication quality	Expectation setting	Q7	Communicating with the chatbot was clear.
		Communication effort	Q10	I had to rephrase my input multiple times for the chatbot to be able to help me.
		Ability to maintain themed discussion	Q15	The chatbot maintained relevant conversation.
		Reference to service	Q16	The chatbot guided me to the relevant service.
		Recognition and facilitation of user's goal and intent	Q24	I find that the chatbot understands what I want and helps me achieve my goal.
		Relevance	Q27	The chatbot provided relevant information as and when I needed it.
		Maxim of quantity	Q29	The chatbot gives me the appropriate amount of information
		Graceful breakdown	Q31	The chatbot could handle situations in which the line of conversation was not clear
		Understandability	Q34	I found the chatbot's responses clear.
		Perceived credibility	Q37	I feel like the chatbot's responses were accurate.
F3	Perceived privacy	Perceived privacy	Q19	The interaction with the chatbot felt secure in terms of privacy.
F4	Perceived speed	Perceived speed	Q42	The chatbot is quick to respond.

Age groups

We determined the optimal 14-item USIC for each age group separately to assess the influence the participant group's characteristics onto the 14-item USIC. Similar to before, we selected for each feature the item with the highest factor loading.

Some different items were selected for the two age groups as compared to the complete participant group's 14-item USIC (see Table 7). However, the two age groups had comparably high factor loadings for the items selected for the complete participant group. The difference in factor loading between the items selected for each age group and the total participant group were negligible (see Appendix C, Table C6 and Table C7). The differences

were slightly larger for the items associated with the features Understandability and Perceived credibility but the complete participant's groups items still provided a good measure for the underlying factor. As such, we advise to use the same 14-item USIC for all age categories and not use distinct compilations or age-related versions of the USIC.

Limitations and recommendations for future research

We consider the proposed 14-item USIC a promising questionnaire due to its compact format which makes it more feasible to implement. We provided preliminary evidence of the 14-item USIC's validity and reliability and recommend further evaluation to continue the standardization process for the reduced USIC as new scales require repeated assessments of its validity and reliability to become a standardized measure (Kyriazos & Stalikas, 2018). We also recommend continued evaluation of the 14-item USIC due to a possible change in context caused by the reduced number of questions. Questions within a scale are not independent. Reducing the USIC's length might therefore affect how individuals answer the remaining questions due to a change in context.

Moreover, the findings indicated that the younger participant group had a slightly different underlying factor structure. A notable difference between the younger and older group is the number of participants with prior experience with chatbots. The majority of the younger group had limited prior experience with chatbots opposed to half of the older group. Earlier research found that the dimensionality of the SUS and UMUX-Lite, that measure user satisfaction with systems, was affected by the individual's level of prior experience (Borsci et al., 2015). Considering the USIC's relation with the UMUX-Lite, the findings indicate that prior experience may influence the USIC. As such, we recommend to conduct further research to explore the influence of users' prior experience with information chatbots on the factor distribution.

The current study evaluated and provided insight into the validity and reliability under a population of individuals between 25-35 and 55-70 years of age. However, we cannot make statements about the USIC's validity and reliability for individuals who do not fall into one of these age groups, such as individuals between 35 and 55 years of age or individuals under the age of 25 or over 70. That said, in previous research the strongest difference in interactive media usage was found between the Millennial and Baby Boomers generations (i.e., individuals similar in age as the groups included here) and Generation X (i.e., individuals between 35-55 years of age) showed an intermediate usage as compared to Millennials and Baby Boomers (Moore, 2012). Taking into account the interactive media usage and the identified similarities in item distribution and factor structure between this study's two age groups, we expect that similar result could be found for individuals between 35 and 55 years of ages. To further increase the generalizability, future studies should include individuals working with chatbots from all age groups.

Furthermore, we recommend to evaluate and optimize the USIC's phrasing. Some participants considered some questions to be ambiguous, or expressed the desire for a "non-applicable" answer category. Participants mainly expressed confusion about item Q17 and to a lesser extent about Q14 and Q40.

- Q17 asks users whether the chatbot uses hyperlinks to guide them to their goals. However, some participants noted that, although the chatbot provided them with hyperlinks, those links did not help them achieve their goal. As such, they were unsure whether they should agree to Q17 because the chatbot did provide links, or whether they should disagree because the hyperlinks provided did not help them achieve their goal.
- For Q14, a couple participants were unsure about what was meant by "context." They wondered what aspects they should take into account when answering the question.

- For Q40, a few participants commented on the ambiguity of “reasonable.” For example, one participant considered the chatbot to answer too fast and selected “don’t agree”, whereas another participant selected “strongly disagree” when she considered the chatbot’s short reaction time to be pleasant.

Wording that can be interpreted in multiple ways should be avoided in scales (Fowler, 2009; Kyriazos & Stalikas, 2018). As such, we recommend to evaluate and optimize the USIC’s phrasing. That said, items Q14, Q17, and Q40 are not included in the 14-item USIC.

Conclusion

The current study contributed to the standardization of the USIC by providing evidence for its validity and reliability under a new population of individuals between 25-35 and 55-70 years old. The findings show that the USIC’s structure is in line with previous studies, it has a strong correlation with the UMUX-Lite, and it has a high internal consistency.

The USIC presents itself as a promising candidate to fulfil the need for a standardized diagnostic scale to measure user satisfaction with information chatbots which was lacking in the literature. The proposed 14-item USIC is especially promising as it is more compact, which makes it more efficient and thus more feasible to implement. The USIC enables researchers and chatbot developers to gain more insight into the user’s satisfaction with various information chatbot and offers the possibility to improve the chatbot in a targeted manner.

References

Abashev, A., Grigoryev, R., Grigorian, K., & Boyko, V. (2017). Programming tools for messenger-based chatbot system organization: Implication for outpatient and translational medicines. *BioNanoScience*, *7*(2), 403–407. https://doi.org/10.1007/s12668-016-0376-9

Artificial Solutions Inc. (2020). *Chatbots: The definitive guide (2020)*. https://www.artificial-solutions.com/chatbots

Balaji, D., & Borsci, S. (2019). *Assessing user satisfaction with information chatbots: A preliminary investigation* [Master's thesis, University of Twente]. http://purl.utwente.nl/essays/79785

Beaudry, J., Consigli, A., Clark, C., & Robinson, K. J. (2019). Getting ready for adult healthcare: Designing a chatbot to coach adolescents with special health needs through the transitions of care. *Journal of Pediatric Nursing*, *49*, 85–91. https://doi.org/10.1016/j.pedn.2019.09.004

Berkman, M. I., & Karahoca, D. (2016). Re-assessing the usability metric for user experience (UMUX) scale. *Journal of Usability Studies*, *11*(3), 89–109.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

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

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations.

Interactions, *25*(5), 38–43. https://doi.org/10.1145/3236669

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. *Usability Evaluation in Industry*, *189*. https://doi.org/10.1002/hbm.20701

Capgemini. (2019). *Smart talk: How organisations and consumers are embracing voice and chat assistants*. https://www.capgemini.com/wp-content/uploads/2019/09/Report_Conversational-Interfaces-1.pdf

Clement, J. (2020). *Number of monthly active Facebook users worldwide as of 1st quarter 2020(in millions)*. Statista. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105–123.

Federici, S., de Filippis, M. L., Mele, M. L., Borsci, S., Bracalenti, M., Gaudino, G., Cocco, A., Amendola, M., & Simonetti, E. (2020). Inside pandora’s box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disability and Rehabilitation: Assistive Technology*, 1–6. https://doi.org/10.1080/17483107.2020.1775313

Field, A. (2009). *Discovering statistics using SPSS (and sex and drugs and rock’n’roll)* (3rd ed.). Sage publications.

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

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

Fowler, F. J. (2009). Designing questions to be good measures. In *Survey Research Methods*

(4th ed.). https://doi.org/10.4135/9781452230184.n6

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. *Proceedings of the International Conference on Information Systems (ICIS)*.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Pearson Education.

International Organization for Standardization. (2018). *Ergonomic of human system interaction - Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11). https://www.iso.org/standard/63500.html

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

Juniper Research. (2019). *Bank cost savings via chatbots to reach $7.3 billion by 2023, as automated customer experience evolves*. https://www.juniperresearch.com/press/press-releases/bank-cost-savings-via-chatbots-reach-7-3bn-2023

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kyriazos, T. A. (2018). Applied psychometrics: The 3-faced construct validation method, a routine for evaluating a factor structure. *Psychology*, *9*(8), 2044–2072. https://doi.org/10.4236/psych.2018.98117

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, *9*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

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

https://doi.org/10.1145/2470654.2481287

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface*. Springer. https://doi.org/10.4135/9781446280409.n3

Moore, M. (2012). Interactive media usage among millennial consumers. *Journal of Consumer Marketing*, *29*(6), 436–444. https://doi.org/10.1108/07363761211259241

Paikari, E., & van der Hoek, A. (2018). A framework for understanding chatbots and their future. *Proceedings of the 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE 2018)*, 13–16.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage Publications.

Plexal. (2018). *Meet Zara: The Zurich UK insurance chatbot co-developed by Plexal member Spixii*. https://www.plexal.com/meet-zara-the-zurich-uk-insurance-chatbot-co-developed-by-plexal-member-spixii/

Radziwill, N., & Benton, M. (2017). Evaluating quality of chatbots and intelligent conversational agents. *ArXiv Preprint ArXiv:1704.04579*.

Research and Markets. (2019). *Chatbot market by component (solutions and services), usage (websites and contact centers), technology, deployment model, application (customer support and personal assistant), organization size, vertical, and region - global forecast to 2024*. https://www.researchandmarkets.com/reports/4858082/chatbot-market-by-component-solutions-and#pos-0

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business*

Research, *57*(2), 184–208. https://doi.org/10.1016/S0148-2963(01)00302-2

Skjuve, M. B., & Brandtzaeg, P. B. (2018). Measuring user experience in chatbots: An approach to interpersonal communication competence. *International Conference on Internet Science*, 113–120. https://doi.org/10.1007/978-3-030-17705-8

Somasundaram, S., Kant, A., Rawat, M., & Maheshwari, P. (2019, February). *The future of chatbots in insurance*. Cognizant. https://www.cognizant.com/whitepapers/the-future-of-chatbots-in-insurance-codex4122.pdf

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM)*, *5*(3), 28–36.

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

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago: University of Chicago Press.

Waldera, L., & Borsci, S. (2019). *Development of a preliminary measurement tool of user satisfaction for information-retrieval chatbots* [Bacherlor's thesis, University of Twente]. http://purl.utwente.nl/essays/79258

Appendices

Appendix A

Table A1.
The 14 chatbot features that Balaji and Borsci (2019) based the USIC on.

	Chatbot feature	Description		Questionnaire item
1	Ease of starting a conversation	How easy it is to start interacting with the chatbot	Q1	It was clear how to start a conversation with the chatbot.
			Q2	It was easy for me to understand how to start the interaction with the chatbot.
			Q3	I find it easy to start a conversation with the chatbot.
2	Accessibility	The ease with which the user can access the chatbot	Q4	The chatbot was easy to access.
			Q5	The chatbot function was easily detectable.
			Q6	It was easy to find the chatbot.
3	Expectation setting	The extent to which the chatbot sets expectations for the interaction with an emphasis on what it can and cannot do	Q7	Communicating with the chatbot was clear.
			Q8	I was immediately made aware of what information the chatbot can give me.
			Q9	It is clear to me early on about what the chatbot can do.
4	Communication effort	The ease with which the chatbot understands a range of user input	Q10	I had to rephrase my input multiple times for the chatbot to be able to help me.
			Q11	I had to pay special attention regarding my phrasing when communicating with the chatbot.
			Q12	It was easy to tell the chatbot what I would like it to do.
5	Ability to maintain themed discussion	The ability of the chatbot to maintain a conversational theme once introduced and keep track of context	Q13	The interaction with the chatbot felt like an ongoing conversation.
			Q14	The chatbot was able to keep track of context.
			Q15	The chatbot maintained relevant conversation.
6	Reference to service	The ability of the chatbot to make references to the relevant service	Q16	The chatbot guided me to the relevant service.
			Q17	The chatbot is using hyperlinks to guide me to my goal.
			Q18	The chatbot was able to make references to the website or service when appropriate.
7	Perceived privacy	The extent to which the user feels the chatbot protects one’s privacy	Q19	The interaction with the chatbot felt secure in terms of privacy.
			Q20	I believe the chatbot informs me of any possible privacy issues.

			Q21	I believe that this chatbot maintains my privacy.
8	Recognition and facilitation of the user's goal and intent	The ability of the chatbot to understand the user's intention and help them accomplish their goal	Q22	I felt that my intentions were understood by the chatbot.
			Q23	The chatbot was able to guide me to my goal.
			Q24	I find that the chatbot understands what I want and helps me achieve my goal.
9	Relevance	The ability of the chatbot to provide information that is relevant and appropriate to the user's request	Q25	The chatbot gave relevant information during the whole conversation.
			Q26	The chatbot is good at providing me with a helpful response at any point of the process.
			Q27	The chatbot provided relevant information as and when I needed it.
10	Maxim of quantity	The ability of the chatbot to respond in an informative way without adding too much information	Q28	The amount of received information was neither too much nor too less.
			Q29	The chatbot gives me the appropriate amount of information.
			Q30	The chatbot only gives me the information I need.
11	Graceful breakdown	The ability of the chatbot to respond appropriately when it encounters a situation it cannot handle	Q31	The chatbot could handle situations in which the line of conversation was not clear.
			Q32	The chatbot explained gracefully when it could not help me.
			Q33	When the chatbot encountered a problem, it responded appropriately.
12	Understandability	The ability of the chatbot to communicate clearly and in an easily understandable manner	Q34	I found the chatbot's responses clear.
			Q35	The chatbot only states understandable answers.
			Q36	The chatbot's responses were easy to understand.
13	Perceived credibility	The extent to which the user believes the chatbot's responses to be correct and reliable	Q37	I feel like the chatbot's responses were accurate.
			Q38	I believe that the chatbot only states reliable information.
			Q39	It appeared that the chatbot provided accurate and reliable information.
14	Perceived speed	The ability of the chatbot to respond timely to user's requests	Q40	The time of the response was reasonable.
			Q41	My waiting time for a response from the chatbot was short.
			Q42	The chatbot is quick to respond.

Note. Adapted from "Assessing User Satisfaction with Information Chatbots: A Preliminary Investigation" by D. Balaji and S. Borsci, 2019, *Master's Thesis, University of Twente*.

Table A2.
The USIC's original wording, its initial and final translation to Dutch and back its translations to English

			Back translation		
	Original English text	Initial translation to Dutch	Translator 1	Translator 2	Final translation to Dutch*
Q1	It was clear how to start a conversation with the chatbot.	Het was duidelijk hoe ik een gesprek met de chatbot kon beginnen.	It was clear how I could start a conversation with the chatbot.	It was immediately clear to me how I could start a conversation with the chatbot.	Het was duidelijk hoe ik een gesprek met de chatbot kon beginnen.
Q2	It was easy for me to understand how to start the interaction with the chatbot.	Het was gemakkelijk te begrijpen hoe ik een gesprek met de chatbot kon beginnen.	It was easy to understand how I could start a conversation with the chatbot.	It was easy to understand how I could start a conversation with the chatbot.	Het was gemakkelijk te begrijpen hoe ik een gesprek met de chatbot kon beginnen.
Q3	I find it easy to start a conversation with the chatbot.	Ik vond het makkelijk om een gesprek met de chatbot te beginnen.	I found it easy to start a conversation with the chatbot.	I found starting a conversation with the chatbot easy.	Ik vond het makkelijk om een gesprek met de chatbot te beginnen.
Q4	The chatbot was easy to access.	De chatbot was makkelijk bereikbaar.	The chatbot was easily accessible.	The chatbot was easily accessible.	De chatbot was makkelijk bereikbaar.
Q5	The chatbot function was easily detectable.	De chatbot functie was makkelijk te ontdekken.	The chatbot function was easy to discover.	The chatbot function was easy to find.	De chatbot functie was makkelijk te ontdekken.
Q6	It was easy to find the chatbot.	Het was makkelijk om de chatbot te vinden.	It was easy to find the chatbot.	Finding the chatbot was easy.	Het was makkelijk om de chatbot te vinden.
Q7	Communicating with the chatbot was clear.	De communicatie met de chatbot was duidelijk.	The communication with the chatbot was clear.	Communication with the chatbot was clear.	De communicatie met de chatbot was duidelijk.
Q8	I was immediately made aware of what information the chatbot can give me.	Ik werd meteen op de hoogte gebracht van de informatie die de chatbot mij kan geven.	I was notified immediately of the information the chatbot could provide.	I was instantly informed about the information that the chatbot has to offer (me).	Ik werd meteen op de hoogte gebracht van de informatie die de chatbot mij kan geven.
Q9	It is clear to me early on about what the chatbot can do.	Het was voor mij snel duidelijk wat de chatbot kan.	It was quickly clear to me what the chatbot can do.	It was immediately clear to me what the chatbot can do.	Het was voor mij **al gauw** duidelijk wat de chatbot kan.
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me.	Ik moest mijn invoer meerdere keren herformuleren voordat de chatbot me kon helpen.	I had to rephrase my entry multiple times before the chatbot could help me.	I had to rephrase my input multiple times before the chatbot was able to help me.	Ik moest mijn invoer meerdere keren herformuleren voordat de chatbot me kon helpen.
Q11	I had to pay special attention regarding my phrasing when communicating with the chatbot.	Ik moest extra goed op mijn formulering letten bij mijn communicatie met de chatbot.	I had to pay close attention to my phrasing during my communication with the chatbot.	I had to formulate extra carefully in my communication with the chatbot.	Ik moest extra goed op mijn formulering letten **tijdens het communiceren** met de chatbot.

			Back translation		
	Original English text	Initial translation to Dutch	Translator 1	Translator 2	Final translation to Dutch*
Q12	It was easy to tell the chatbot what I would like it to do.	Het was makkelijk om de chatbot te vertellen wat ik wilde dat het deed.	It was easy to tell the chatbot what I wanted it to do.	Communicating my requests to the chatbot was easy.	Het was makkelijk om de chatbot te vertellen wat ik wilde dat het deed.
Q13	The interaction with the chatbot felt like an ongoing conversation.	De interactie met de chatbot voelde als een lopend gesprek.	The interaction with the chatbot felt as a fluent conversation.	The interaction with the chatbot felt like a fluid conversation.	De interactie met de chatbot voelde als een lopend gesprek.
Q14	The chatbot was able to keep track of context.	De chatbot hield de context in het oog.	The chatbot paid attention to the context.	The chatbot was attentive/responsive to the context of what was (being) said.	De chatbot hield de context in het oog.
Q15	The chatbot maintained relevant conversation.	Het gesprek wat de chatbot onderhield was relevant.	The conversation held by the chatbot was relevant.	The chatbot held relevant conversation.	Het gesprek dat de chatbot **voerde** was relevant.
Q16	The chatbot guided me to the relevant service.	De chatbot leidde me naar de relevante service.	The chatbot guided me to the relevant service.	The chatbot directed me to the relevant service.	De chatbot leidde me naar de relevante service.
Q17	The chatbot is using hyperlinks to guide me to my goal.	De chatbot gebruikte hyperlinks om me naar mijn doel te leiden.	The chatbot used hyperlinks to guide me to my goal.	The chatbot made use of hyperlinks to direct me to my goal.	De chatbot gebruikte hyperlinks om me naar mijn doel te leiden.
Q18	The chatbot was able to make references to the website or service when appropriate.	De chatbot kon naar de website of dienst verwijzen wanneer dat nodig was.	The chatbot could direct me to the website or service if needed.	The chatbot was able to direct me to relevant websites or services when needed.	De chatbot kon me **verwijzen** naar de website of **een** dienst **wanneer nodig**.
Q19	The interaction with the chatbot felt secure in terms of privacy.	De interactie met de chatbot voelde veilig in relatie tot privacy.	The interaction with the chatbot felt safe in relation to privacy	The interaction with the chatbot felt safe in terms of privacy.	De interactie met de chatbot voelde veilig **met betrekking** tot privacy.
Q20	I believe the chatbot informs me of any possible privacy issues.	Ik denk dat de chatbot me inlicht over mogelijke privacy problemen.	I think the chatbot informs me about possible privacy issues.	I believe that the chatbot informs me of/about possible privacy issues.	Ik denk dat de chatbot me inlicht over mogelijke privacy problemen.
Q21	I believe that this chatbot maintains my privacy.	Ik denk dat de chatbot mijn privacy beschermd.	I think the chatbot protects my privacy.	I believe the chatbot safeguards my privacy.	Ik denk dat de chatbot mijn privacy **waarborgt**.
Q22	I felt that my intentions were understood by the chatbot.	Ik had het gevoel dat mijn intenties werden begrepen door de chatbot.	I had the feeling that my intentions were understood by the chatbot.	I felt like my intentions were understood by the chatbot.	Ik had het gevoel dat mijn intenties werden begrepen door de chatbot.
Q23	The chatbot was able to guide me to my goal.	De chatbot begeleidde mij naar mijn doel.	The chatbot guided me to my goal.	The chatbot directed me to my goal.	De chatbot begeleidde mij naar mijn doel.

			Back translation		
	Original English text	Initial translation to Dutch	Translator 1	Translator 2	Final translation to Dutch*
Q24	I find that the chatbot understands what I want and helps me achieve my goal.	Ik denk dat de chatbot begrijpt wat ik wil en helpt mijn doel te bereiken.	I think the chatbot understands what I want and helps to reach my goal.	I believe the chatbot understands my needs and helps me in achieving my goal.	Ik denk dat de chatbot begrijpt wat ik wil en helpt me mijn doel te bereiken.
Q25	The chatbot gave relevant information during the whole conversation.	De chatbot gaf tijdens het gehele gesprek relevante informatie.	The chatbot provided relevant information during the entire conversation.	The chatbot provided relevant information during the entire conversation.	De chatbot gaf tijdens het gehele gesprek relevante informatie.
Q26	The chatbot is good at providing me with a helpful response at any point of the process.	De chatbot gaf behulpzame reacties tijdens het gehele gesprek.	The chatbot provided helpful responses during the entire conversation.	The chatbot's reactions were helpful during the entire conversation.	De chatbot gaf behulpzame reacties **op elk moment in het proces**.
Q27	The chatbot provided relevant information as and when I needed it.	De chatbot gaf relevante informatie wanneer ik die nodig had.	The chatbot provided relevant information whenever I needed that.	The chatbot provided relevant information when needed.	De chatbot gaf relevante informatie wanneer ik die nodig had.
Q28	The amount of received information was neither too much nor too less.	De hoeveelheid informatie die ik ontving was niet te veel en niet te weinig.	The amount of information I received was not too much nor too little.	The amount of information I received was not too much and not too little.	De hoeveelheid informatie die ik ontving was niet te veel en niet te weinig.
Q29	The chatbot gives me the appropriate amount of information.	De chatbot gaf me de juiste hoeveelheid informatie.	The chatbot provided me the right amount of information.	The amount of information I received was just right.	De chatbot gaf me de juiste hoeveelheid informatie.
Q30	The chatbot only gives me the information I need.	De chatbot gaf me alleen de informatie die ik nodig had.	The chatbot only provided me with the information I needed.	The chatbot only provided the information I needed.	De chatbot gaf me alleen de informatie die ik nodig had.
Q31	The chatbot could handle situations in which the line of conversation was not clear.	De chatbot kon omgaan met situaties waarin de rode draad van het gesprek niet duidelijk was.	The chatbot could handle situations in which the red line of the conversation was not clear.	The chatbot could adequately deal with situations where the direction of the conversation was unclear.	De chatbot kon omgaan met situaties waarin de rode draad van het gesprek niet duidelijk was.
Q32	The chatbot explained gracefully when it could not help me.	De chatbot vertelde me op een beleefde manier wanneer het me niet kon helpen.	The chatbot told me in a polite manner when it could not help me.	The chatbot informed me politely when it could not be of assistance (to me).	De chatbot vertelde me op een **vriendelijke** manier wanneer het me niet kon helpen.
Q33	When the chatbot encountered a problem, it responded appropriately.	Als de chatbot op een probleem stuitte, reageerde hij op gepaste wijze.	If the chatbot came across an issue, he responded in an appropriate manner.	The chatbot reacted appropriately whenever it encountered a problem.	Als de chatbot op een probleem stuitte, reageerde het op gepaste wijze.
Q34	I found the chatbot's responses clear.	Ik vond de antwoorden van de chatbot duidelijk.	I considered the answers of the chatbot clear.	The chatbot's answers were clear.	Ik vond de antwoorden van de chatbot duidelijk.

			Back translation		
	Original English text	Initial translation to Dutch	Translator 1	Translator 2	Final translation to Dutch*
Q35	The chatbot only states understandable answers.	De chatbot gaf alleen begrijpelijke antwoorden.	The chatbot only provided understandable answers.	The chatbot only answered comprehensively.	De chatbot gaf alleen begrijpelijke antwoorden.
Q36	The chatbot's responses were easy to understand.	De antwoorden van de chatbot waren gemakkelijk te begrijpen.	The answers of the chatbot were easy to understand.	The answers given by the chatbot were easy to understand.	De antwoorden van de chatbot waren gemakkelijk te begrijpen.
Q37	I feel like the chatbot's responses were accurate.	Ik had het gevoel dat de antwoorden van de chatbot juist waren.	I had the feeling that the answers of the chatbot were right.	I felt that the chatbot's answers were accurate.	Ik had het gevoel dat de antwoorden van de chatbot **klopten**.
Q38	I believe that the chatbot only states reliable information.	Ik denk dat de chatbot alleen betrouwbare informatie geeft.	I think the chatbot only provides reliable information.	I believe that the chatbot only provides dependable information.	Ik denk dat de chatbot alleen betrouwbare informatie geeft.
Q39	It appeared that the chatbot provided accurate and reliable information.	De informatie die de chatbot gaf leek betrouwbaar en juist.	The information the chatbot provided seemed reliable and correct.	The information provided by the chatbot appeared trustworthy and correct.	De informatie die de chatbot gaf leek betrouwbaar en juist.
Q40	The time of the response was reasonable.	De reactietijd van de chatbot was acceptabel.	The response time of the chatbot was acceptable.	The chatbot's response time was acceptable.	De reactietijd van de chatbot was **redelijk**.
Q41	My waiting time for a response from the chatbot was short.	Ik hoefde kort te wachten op een antwoord van de chatbot.	I had to wait shortly for an answer from the chatbot.	I had to wait a short time for the chatbot to reply.	Ik hoefde kort te wachten op een antwoord van de chatbot.
Q42	The chatbot is quick to respond.	De chatbot reageerde snel.	The chatbot responded quickly.	The chatbot responded quickly.	De chatbot reageerde snel.

* Revisions after back translations in **bold**

Table A3.
The UMUX-Lite's original wording, its initial and final translation to Dutch and back its translations to English

	Original English text	Initial translation to Dutch	Back translation		Final translation to Dutch
			Translator 1	Translator 2	
Q1	This system's capabilities meet my requirements.	De mogelijkheden die dit systeem biedt voldoen aan mijn eisen.	The possibilities this system offers meet my expectations.	The possibilities provided by this system meet my expectations.	De mogelijkheden die dit systeem biedt voldoen aan mijn eisen.
Q2	This system is easy to use.	Dit systeem is makkelijk te gebruiken.	This system is easy to use.	This system is easy to use.	Dit systeem is makkelijk te gebruiken.

Appendix B

Informed consent form (English)

Information sheet for the research study:

Questionnaire on user satisfaction of chatbots

We invited you to participate in this research study about a questionnaire that assess the user's satisfaction of chatbots. This research is led by Imke Silderhuis.

Please read this consent form carefully and ask the researcher about anything that is unclear.

Research goal

The goal of this research study is to evaluate the questionnaire about the user's satisfaction of chatbots. The collected data will be used for educative and scientific purposes (e.g., a publication).

What will we do?

In this study, the researcher will work with you using an audio- and screen-connection. We will first ask you to fill out some background questions (e.g., age, gender, educational background, experience with chatbots). Then, you will interact with and perform tasks with five chatbots, during which we will ask you think aloud. After performing a task with a chatbot, we will ask you to complete a questionnaire about the associated chatbot.

After you filled out the background questions, we will start an audio- and computer screen recording to retrieve information on how users perceive the chatbots. We might also take some notes.

In summary, we will collect information by:

- Have you fill out questionnaires on the computer.
- Recording of audio and computer screen (during the chatbot tasks and filling out the related questionnaires).
- Observation.

Confidentiality of data

We do everything we can to protect your privacy as well as possible. We will not associate your name with the data, but instead will use pseudonyms (e.g., "participant 1") to anonymise your data.

No confidential information or personal data is released from or about you that you could be recognized from. Audio recordings will only by accessed the research group and will not be released. The research data could only made available in anonymous form, if necessary (for example for a check on scientific integrity) and to persons outside the research group.

The audio recordings, forms and other documents that are made or collected for this study are stored at a secure location at the University of Twente and on the secured (encrypted) data carriers of the researchers. The research data is stored for a period of 10 years. The data will be deleted after this period at the latest.

Potential risks and discomforts

There are no risks associated with your participation in this study.

Voluntary participation

Participation in this study is completely voluntary. You withdraw from the study at any time, or refuse that your data may be used for the study, without stating why. You don't have to answer any questions that you do not want to answer. Stopping participation or not answering questions does not result in any negative consequences for you.

If, during the study, you decide to stop participating, the information that you have provided up to that point will be used, unless you state otherwise.

Ethics approval

This research study was assessed and approved by the ethics committee of the Faculty of Behavioral Management and Social Sciences (BMS).

For objections regarding the study, you can contact the Secretary of the Ethics Committee of the faculty of Behavioral, Management and Social Sciences at the University of Twente via ethicscommittee-bms@utwente.nl. This research study is being conducted as part of the University of Twente, Behavioral Faculty, Management and Social Sciences.

If you have specific questions about the handling of personal data, you can also address this to the Data Protection Officer of the UT by sending an email to dpo@utwente.nl.

Finally, you have the right to make a request to inspect, change, delete or modify your data with the Researcher.

Contact details

Principal Researcher
Imke Silderhuis

Co-Investigator
Dr. Simone Borsci

Consent form for the research study:

Questionnaire on user satisfaction of chatbots

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that I can refuse to answer questions, without having to give a reason.	○	○
I understand that taking part in the study involves performing tasks with chatbots and filling out questionnaires during which an audio- and screen-recording will be made.	○	○
Use of the information in the study		
I understand that personal information collected about me that can identify me, such as my name or the audio-recording, will not be shared beyond the study team.	○	○
I understand that information I provide will be used for educational and scientific purposes (e.g., publication).	○	○
Consent to be recorded		
I agree to a screen-recording and to be audio recorded.	○	○
Yes, I am 18 years old or older, read the information sheet and am voluntarily participating in this study.	○	○

Informed consent form (Dutch)

Informatieblad van het onderzoek:
‘Vragenlijst over de gebruikerstevredenheid van chatbots’

We hebben u uitgenodigd om mee te doen aan een onderzoek over een vragenlijst die gebruikerstevredenheid van chatbots meet. Dit onderzoek wordt geleid door Imke Silderhuis.

Leest u alstublieft dit informatieblad en toestemmingsformulier zorgvuldig door en vraag de onderzoeker als iets onduidelijk is.

Doel van het onderzoek

Het doel van dit onderzoek is het evalueren van een vragenlijst over gebruikerstevredenheid van chatbots. De onderzoeksgegevens van dit onderzoek zullen gebruikt worden voor educatieve en wetenschappelijke doeleinden (bijvoorbeeld een publicatie).

Hoe gaan we te werk?

In dit onderzoek werkt de onderzoeker met u via een audio- en beeldconnectie. We zullen u eerst vragen een aantal achtergrondvragen in te vullen (bijvoorbeeld uw leeftijd, geslacht, uw scholing en ervaring met chatbots). Daarna, zult u een taak uit te voeren met vijf chatbots, waarbij we u zullen vragen hard op te denken. Na het uitvoeren van de taak zullen we u vragen een vragenlijst in te vullen over de betreffende chatbot.

Nadat u de vragenlijst met achtergrondvragen heeft ingevuld, starten we een opname van het geluid en het computerbeeld. We doen dit om informatie te verzamelen over hoe gebruikers de chatbot ervaren. Ook zullen we notities maken.

Samenvattend, zullen we informatie verzamelen door:

- U een aantal vragenlijsten op de computer in te laten vullen.
- Opname van het geluid en computerbeeld (tijdens de chatbot taken en het invullen van de gerelateerde vragenlijst).
- Observatie.

Vertrouwelijkheid van gegevens

Wij doen er alles aan uw privacy zo goed mogelijk te beschermen. We zullen uw naam niet aan de onderzoeksgegevens koppelen, maar maken gebruik maken van pseudoniemen (bijvoorbeeld ‘participant 1’) om uw data te anonimiseren.

Er wordt geen vertrouwelijke informatie of persoonsgegevens van of over u naar buiten gebracht, waardoor iemand u kan herkennen. Geluidsopnames zijn alleen toegankelijk voor de onderzoeksgroep en worden niet vrijgegeven. De onderzoeksgegevens worden indien nodig (bijvoorbeeld voor een controle op wetenschappelijke integriteit) en alleen in anonieme vorm ter beschikking gesteld aan personen buiten de onderzoeksgroep.

De geluidsopnamen, formulieren en andere documenten die in het kader van deze studie worden gemaakt of verzameld, worden opgeslagen op een beveiligde locatie bij de Universiteit Twente en op de beveiligde (versleutelde) gegevensdragers van de onderzoekers.

De onderzoeksgegevens worden bewaard voor een periode van 10 jaar. De gegevens worden uiterlijk verwijderd na deze termijn.

Potentiële risico's en ongemakken

Er zijn geen risico's verbonden aan uw deelname aan deze studie.

Vrijwilligheid

Deelname aan dit onderzoek is geheel vrijwillig. U kunt op elk moment stoppen, of weigeren dat uw gegevens voor het onderzoek worden gebruikt, zonder dat u hiervoor een reden hoeft te geven. U hoeft geen vragen te beantwoorden die u niet wilt beantwoorden. Het stopzetten van deelname of niet beantwoorden van vragen heeft geen nadelige gevolgen voor u.

Als u tijdens het onderzoek besluit om te stoppen, zullen de gegevens die u al hebt verstrekt tot dat moment in het onderzoek gebruikt worden, tenzij u iets anders aangeeft.

Ethische goedkeuring

Dit onderzoek is beoordeeld en goedgekeurd door de ethische commissie van de faculteit Behavioural Management and Social sciences (BMS).

Voor bezwaren met betrekking tot de opzet en of uitvoering van het onderzoek kunt u contact opnemen met de Secretaris van de Ethische Commissie van de faculteit Behavioural, Management and Social Sciences op de Universiteit Twente via ethicscommittee-bms@utwente.nl. Dit onderzoek wordt uitgevoerd vanuit de Universiteit Twente, faculteit Behavioural, Management and Social Sciences. Indien u specifieke vragen hebt over de omgang met persoonsgegevens kun u deze ook richten aan de Functionaris Gegevensbescherming van de UT door een mail te sturen naar dpo@utwente.nl.

Tot slot heeft u het recht een verzoek te doen tot inzage, wijziging, verwijdering of aanpassing van uw gegevens bij de Onderzoeksleider.

Contact gegevens

Hoofdonderzoeker
Imke Silderhuis

Co-onderzoeker
Dr. Simone Borsci

Toestemmingsformulier van het onderzoek: ‘Vragenlijst over de gebruikerstevredenheid van chatbots’

U ONTVANGT EEN KOPIE VAN DIT TOESTEMMINGSFORMULIER

Kruis a.u.b. de voor u juiste cirkels aan.	**Ja**	**Nee**
Mee doen aan het onderzoek		
Ik heb het informatieblad van gelezen en begrepen, of deze is aan mij voorgelezen. Ik heb vragen kunnen stellen over het onderzoek en mijn vragen zijn naar tevredenheid beantwoord.	○	○
Ik doe vrijwillig mee aan dit onderzoek. Ik begrijp dat ik me op elk moment kan terugtrekken uit het onderzoek, zonder een reden op te geven.	○	○
Ik begrijp dat ik kan weigeren om vragen te beantwoorden, zonder een reden op te geven.	○	○
Ik begrijp dat ik tijdens dit onderzoek taken uitvoer met chatbots en vragenlijst hierover invul, en hiervan een geluidsopname en computerbeeldopname gemaakt wordt.	○	○
Informatiegebruik		
Ik begrijp dat de verzamelde persoonsgegevens die mij kunnen identificeren, zoals mijn naam of de geluidsopname, niet gedeeld worden met personen buiten het onderzoeksteam.	○	○
Ik begrijp dat de informatie die ik verstrek zal worden gebruikt voor educatieve en wetenschappelijke doeleinden (bijvoorbeeld voor een publicatie).	○	○
Toestemming voor opnames		
Ik geef toestemming om een geluidsopname en opname van computerbeeld te maken.	○	○
Ja, ik ben 18 jaar of ouder, heb het informatieblad gelezen en doe vrijwillig mee aan dit onderzoek.	○	○

Session Script (English)

<Participant and researcher will set up a connection using Whereby>

"Hi. My name is Imke. Welcome and thank you for taking the time to participate in today's study. Are you ready to start and have me explain what we are going to do?

<Check if participant is ready to start>

"Great. Before we start, can you silence or switch-off your phone for the duration of the session?

Also, please let me know if my microphone might encounter any issues

For this research, we are evaluating a questionnaire to capture the user's satisfaction for chatbots. Today, I will ask you to work with five chatbots. For each chatbot, we have a brief scenario and one or two tasks. After every task, I will ask you to fill out a questionnaire about your experience with the chatbot. The questionnaire has 42 questions.

Please don't feel nervous or under any kind of pressure. It is not a test of how well you interact with the chatbots. Rather, we are interested in your honest feedback on the chatbots, which you can give by filling out the questionnaire. The session is scheduled to last an hour to 1.5 hours. Do you have any questions at this point?"."

"First, I will send you a link to start the research. I would like you to open the link in a new browser tab."

<Participant opens Qualtrics>

"I have an informed consent form for you. I would like you to read the form. Please let me know if you have any questions. I like to point out that we will make a recording of the audio and screen- today for data-analysis purposes. I will let you know once I will start it. Please let me know if you are not comfortable with this.

If you are ok with everything the form notes, please tick the boxes below the form."

<Have participants read the informed consent form and tick the boxes>

"Before we start off with the chatbot tasks, I have a couple of questions for you regarding your background. Could you please fill these out?"

<Participant fills out background questionnaire.>

"Ok, I would like to ask you to share your screen with me. Please close any other windows on your computer, if you don't want me to see those."

<Participant shares screen>

"Then we will start now with the first task. I will start the screen- and audio-recording now."

<Start audio and screen recording>

"Today, you will work on a task with five chatbots. While you are working on the chatbot task, I will like to share your thoughts with me and tell me what you do and see. The chatbot can send you a link, you can click on these if you like, but I will like to ask to not go much further into the website than that particular page.

The last two out of five chatbots are English. I would like to ask you to talk English with these chatbots. If you have any difficulty with the language, you can ask me for help.

If anything is unclear, you can ask me. However, I may not be able to answer all questions to not influence the research.

Let me know once you achieved the task or if you feel the task is not achievable. We can then move on to the questionnaire.

On the next page, you will see a link to a website on the screen. I would like you to open a new browser tab and copy the link to access the website. You can then continue to the next page in the questionnaire and you will receive the task.

Are you ready to begin?"

<Participant works on chatbot tasks and fills out questionnaires>

That completes all of the planned activities for today. Do you have any questions or comments?

If you know someone who will participate in today's study, I will like to ask you to not discuss the study.

Thank you very much for participating in this study. Your participation is very valuable to us.

Session Script (Dutch)

<Participant en onderzoeker zetten internet connectie op via Whereby>

"Hallo. Mijn naam is Imke. Welkom en bedankt dat u mee wilt doen aan dit onderzoek vandaag. Bent u klaar om te beginnen en zal ik uitleggen wat we vandaag gaan doen?"

<Controleren of participant klaar is om te beginnen>

"Super. Voordat we beginnen, kunt u uw mobiele telefoon op stil of uitzetten voor de sessie?

Als het geluid van mijn microfoon niet goed werkt, laat het me dan weten.

In dit onderzoek evalueren we een vragenlijst die de gebruikerstevredenheid van chatbots meet. Vandaag zal ik u vragen om met vijf chatbots te werken. Voor elke chatbot krijgt u een korte situatieschets en 1 of 2 taken. Na elke taak zal ik u vragen een vragenlijst in te vullen over uw ervaring met de chatbot. Deze vragenlijst heeft 42 vragen.

Voelt u alsjeblieft niet nerveus of onder druk gezet. Dit onderzoek is geen test van hoe goed u met de chatbots omgaat. We zijn geïnteresseerd in uw eerlijke feedback op de chatbots, die u kunt geven door het invullen van de vragenlijst. De gehele sessie duurt een uur tot anderhalf uur. Heb je op dit moment nog vragen?".

"Dan stuur ik u eerst een link sturen om het onderzoek te openen. Deze mag u in een nieuw internet tabblad openen. "

<Participant opent Qualtrics>

"Om te beginnen, heb ik een geïnformeerd toestemmingsformulier voor u. Wilt u dit alstublieft lezen? Ik wil u erop wijzen dat we vandaag een opname maken van de audio en het scherm voor data-analyse. Ik laat het u weten zodra ik deze opname begin. Laat het me alstublieft weten als u zich hier niet prettig bij voelt. En als u vragen heeft over het formulier, hoor ik het graag."

"Als u het eens bent met alles wat er op het formulier staat, vink dan de vakjes aan onder het formulier."

<Participant leest het geïnformeerde toestemmingsformulier en vinkt de vakjes aan>

"Voordat we beginnen met de chatbot-taken, heb ik een aantal vragen voor u over uw achtergrond. Wilt u deze alstublieft invullen?

<Participant vult achtergrondvragenlijst in.>

En dan uw scherm met mij delen? "Ok, dan wil ik u vragen uw scherm met mij te delen. Wilt u eventuele andere schermen sluiten, die u niet wilt dat ik zie.

<Participant deelt scherm>

Dan we beginnen nu met de eerste opdracht. Ik begin nu met de scherm- en audio-opname en geef een kleine toelichting."

<Start-audio en schermopname>

"Vandaag zul je met vijf chatbots een taak doen. Terwijl u aan de chatbot-taak werkt, wil ik u vragen om mij mee te nemen in uw denkproces en hardop na te denken over wat u doet en ziet. Het kan zijn dat de chatbot linkjes stuurt, u kunt hier op klikken maar ik wil u vragen niet verder de website in te gaan.

De laatste twee chatbots van de vijf zijn Engels. Ik wil u vragen u om in het Engels te praten met deze chatbots. Als u moeite hebt met de taal, kun u mij om hulp vragen.

Als er verder iets onduidelijk is, kunt u het mij vragen. Het kan echter zijn dat ik niet al uw vragen kan beantwoorden om het onderzoek niet te beïnvloeden.

Als u denkt de taak volbracht te hebben of denkt dat de taak niet te volbrengen is, mag u het aan mij laten weten. We kunnen dan verder gaan met de vragenlijst.

Op het volgende scherm ziet u een link naar een website. Wilt u de link naar een nieuw internet tabblad te kopiëren? U kunt daarna verder gaan naar de volgende pagina voor de chatbot taak.

Bent u klaar om te beginnen?

<Participant werkt aan chatbot-taken en vult vragenlijsten in>

Dat waren alle activiteiten voor vandaag. Heeft u vragen of opmerkingen?

Als u iemand kent die nog mee gaat doen, dan wil ik u vragen niet te vertellen over het onderzoek.

Hartelijk dank voor uw deelname aan dit onderzoek. Uw deelname is erg waardevol voor ons.

Table B1.
Participant demographics questionnaire

	English	Dutch
Age	What is your age?	Wat is uw leeftijd?
Gender	What is your gender? • Male • Female • Other • Prefer not to say	Wat is uw geslacht? • Man • Vrouw • Anders • Zeg ik liever niet
Dutch proficiency	How well do you read, write, and understand Dutch? How fluent are you in Dutch? • Excellent/ Native • Good • Moderate • Basic knowledge • None	Hoe goed leest, schrijft en begrijpt u Nederlands? Hoe vloeiend bent u in het Nederlands? • Uitstekend/ Moedertaal • Goed • Matig • Basiskennis • Niet
English proficiency	How well do you read, write, and understand English? How fluent are you in English? • Excellent/ Native • Good • Moderate • Basic knowledge • None	Hoe goed leest, schrijft en begrijpt u Engels? Hoe vloeiend bent u in het Engels? • Uitstekend/ Moedertaal • Goed • Matig • Basiskennis • Niet
Education	What is your highest level of completed education? • Primary school • High school • MBO degree • HBO-bachelor, WO-bachelor • HBO-master, WO-master, PhD	Wat is uw hoogst behaalde opleidingsniveau? • Basisonderwijs • Middelbare school • MBO diploma • HBO-bachelor, WO-bachelor • HBO-master, WO-master, PhD
Familiarity chatbots	How familiar are you with chatbots and/ or other conversational interfaces? • Extremely familiar • Very familiar • Moderately familiar • Slightly familiar • Not familiar at all	Hoe bekend bent u met chatbots en/ of andere gespreksinterfaces? • Uiterst bekend • Erg bekend • Enigszins bekend • Beetje bekend • Niet bekend
Prior usage chatbots	Have you used a chatbot before? • Definitely yes • Probably • Unsure • Probably not • Definitely not	Heeft u eerder een chatbot gebruikt? • Zeker ja • Waarschijnlijk • Niet zeker • Waarschijnlijk niet • Zeker niet
Frequency using chatbot	How often do you use chatbots? • Daily • 4-6 times a week • 2-3 times a week • Once a week • Rarely • Never	Hoe vaak gebruikt u chatbots? • Dagelijks • 4-6 keer per week • 2-3 keer per week • Een keer per week • Zelden • Nooit

Table B2.
Included chatbots and associated URL links

Chatbot	URL link
English chatbots	
Absolut	https://www.absolut.com/en/
Australian Taxation Office (ATO)	https://www.ato.gov.au/
HSBC UK	https://www.hsbc.co.uk/
United States Citizenship and Immigration Services (USCIS)	https://www.uscis.gov/
Dutch chatbots	
Amsterdam Medisch Centrum	https://www.amc.nl/
A.S.R.	https://www.asr.nl/
Bol.com	https://www.bol.com/nl/
KPN	https://www.kpn.com/
Oxxio	https://www.oxxio.nl/
Vattenfall	https://www.vattenfall.nl/

Table B3.
Included chatbots and associated task prompts in English and Dutch

Chatbot	English task prompt	Dutch task prompt
	English chatbots	
Absolut	You want to buy a bottle of Absolut vodka to share with your friends for the evening. One of your friends cannot consume gluten. You want to use Absolut's chatbot to find out if Absolut Lime contains gluten or not.	Je wilt een fles Absolut wodka kopen om te delen met je vrienden 's avonds. Een van je vrienden mag geen gluten innemen. Je wilt de Absolut chatbot gebruiken om te weten te komen of Absolut Lime wodka gluten bevat of niet. *Let op: dit is een Engelse chatbot. Schrijf je vraag in het Engels.*
ATO	You moved to Australia from the Netherlands recently. You want to know when the deadline is to lodge/submit your tax return using ATO's chatbot to find out.	Je bent recentelijk vanuit Nederland naar Australië verhuisd. Je wilt weten wanneer de deadline is om je belastingsaangifte te doen en gebruikt de ATO chatbot om meer te weten te komen. *Let op: dit is een Engelse chatbot. Schrijf je vraag in het Engels.*
HSBC UK	You live in the Netherlands but are travelling to Turkey for 2 weeks. During your travel, you would like to be able to use your HSBC credit card overseas at payment terminals and ATMS. You want to use HSBC's chatbot to find out the relevant procedure.	Je woont in Nederland en reist voor twee weken naar Turkije. Tijdens je reis wil je graag je HSBC credit card kunnen gebruiken bij betaal- en geldautomaten. Je wilt de HSBC chatbot gebruiken om de relevante procedure te weten te komen. *Let op: dit is een Engelse chatbot. Schrijf je vraag in het Engels.*
United States Citizenship and Immigration Services (USCIS)	You are a US citizen living abroad and want to vote in the upcoming federal elections. You want to use the USCIS chatbot to find out how.	Je bent een Amerikaanse staatsburger die in het buitenland woont. Je wilt stemmen bij de komende federale verkiezingen. Je wilt de USCIS chatbot gebruiken om uit te vinden hoe je dat kunt doen. *Let op: dit is een Engelse chatbot. Schrijf je vraag in het Engels.*
	Dutch chatbots	
Amsterdam Medisch Centrum	You need to get your blood tested at the Amsterdam Medical Center (AMC). You want to use the chatbot to find out where in the hospital you need to be and what the procedure is for blood sampling.	Je moet bloed laten prikken in het Amsterdam Medisch Centrum (AMC) voor een onderzoek. Je wilt de chatbot gebruiken om erachter te komen waar je in het ziekenhuis moet zijn, en wat de procedure is bij bloedprikken.
A.S.R.	Your motorbike has been hit while you were parked at a gas station. You can't continue driving. You are insured with ASR and visit the website to report the damage and see if you can get a replacement vehicle.	Je motor is aangereden terwijl je geparkeerd stond bij een benzinestation. Je kunt niet meer verder rijden. Je bent verzekerd bij ASR en gebruikt de chatbot om de schade te melden en om te kijken of je vervangend vervoer kunt krijgen.
Bol.com	You forgot to buy a present for a friend who is celebrating her birthday tonight, and you want to buy a 10 euro Bol.com gift card. You want to use the Bol.com	Je bent vergeten een cadeau te kopen voor een vriendin die vanavond haar verjaardag viert en je wilt nog snel een Bol.com cadeaukaart van 10 euro kopen.

	chatbot to find out in which shop you can buy the gift card and what is the lowest amount you can put on a gift card.	Je wilt de Bol.com chatbot gebruiken om erachter te komen in welke winkel je de cadeaukaart kunt kopen en wat het laagste bedrag is wat je op een cadeaukaart kunt zetten.
KPN	You are a KPN customer and have a prepaid SIM card for your mobile phone. You need new prepaid credit and you want to use the chatbot to find out how long prepaid credit is valid after purchase. Additionally, you have a new account number that you want to pass on to KPN. You want to use the KPN chatbot to find out how you can change your account number.	Je bent klant bij KPN en hebt een prepaid simkaart voor je mobiele telefoon. Je hebt nieuw prepaid tegoed nodig. Je wilt met behulp van de KPN chatbot te weten komen hoe lang prepaid tegoed geldig is na aankoop. Daarnaast heb je een nieuw rekeningnummer dat je door wilt geven aan KPN. Je wilt door middel van de KPN chatbot te weten komen hoe je jouw rekeningnummer kunt wijzigen.
Oxxio	You're considering switching to the Oxxio's green energy. However, the contract with your current energy supplier has not yet ended, and your energy supplier will impose a cancellation penalty if you switch suppliers before the end date. You want to use the chatbot to find out whether Oxxio will pay this fine for you if you switch to Oxxio.	Je overweegt om over te stappen naar de duurzame stroom van Oxxio. Het contract bij je huidige energieleverancier is echter nog niet afgelopen, en je energieleverancier rekent een opzegboete als je voor de einddatum overstapt. Je wilt er met behulp van de chatbot achter komen of Oxxio deze boete betaalt voor jou als naar Oxxio overstapt.
Vattenfall	You are a Vattenfall customer and receive monthly 'exclusive points', which you can donate to charity, among other things. You want to ask the chatbot which charities you can donate these 'exclusive points' to.	Je bent klant bij Vattenfall en krijgt maandelijks 'exclusief punten', die je o.a. kunt doneren aan het goede doel. Je wilt de chatbot vragen aan welke goede doelen je deze 'exclusief punten' kunt doneren.

Appendix C

Table C1.
Participant demographics

Characteristics	Total	Age category 25-35 years old	Age category 55-70 years old
Age (years)	Average = 45.55, SD = 17.21	Average = 28.8, SD = 2.7	Average = 62.3, SD = 3.89
Gender	Female = 36, Male = 24	Female = 18, Male = 12	Female = 18, Male = 12
Dutch proficiency	Excellent/ native = 56 Good = 4 Moderate = 0 Basic knowledge = 0	Excellent/ native = 29 Good = 1 Moderate = 0 Basic knowledge = 0	Excellent/ native = 27 Good = 3 Moderate = 0 Basic knowledge = 0
English proficiency	Excellent/ native = 7 Good = 40 Moderate = 12 Basic knowledge = 1	Excellent/ native = 5 Good = 24 Moderate = 1 Basic knowledge = 0	Excellent/ native = 2 Good = 16 Moderate = 11 Basic knowledge = 1
Education level**	HBO-master, WO-master, PhD = 26 HBO-bachelor, WO-bachelor = 22 Intermediate vocational education/ MBO diploma = 10 High school = 2	HBO-master, WO-master, PhD = 21 HBO-bachelor, WO-bachelor = 5 Intermediate vocational education/ MBO diploma = 2 High school = 2	HBO-master, WO-master, PhD = 5 HBO-bachelor, WO-bachelor = 17 Intermediate vocational education/ MBO diploma = 8 High school = 0
Familiarity chatbots	Very familiar = 11 Moderately familiar = 24 Slightly familiar = 18 Not familiar = 7	Very familiar = 8 Moderately familiar = 13 Slightly familiar = 9 Not familiar = 0	Very familiar = 3 Moderately familiar = 11 Slightly familiar = 9 Not familiar = 7
Prior usage chatbots	Definitely yes = 34 Probably = 11 Unsure = 1 Probably not = 7 Definitely not = 7	Definitely yes = 23 Probably = 6 Unsure = 0 Probably not = 1 Definitely not = 0	Definitely yes = 11 Probably = 5 Unsure = 1 Probably not = 6 Definitely not = 7
Frequency using chatbots	Daily = 1 4 – 6 times a week = 0 2 – 3 times a week = 1 Once a week = 5 Rarely = 27 Never = 0 Previously indicated (probably) not used a chatbot before = 14	Daily = 1 4 – 6 times a week = 0 2 – 3 times a week = 1 Once a week = 1 Rarely = 26 Never = 0 Previously indicated (probably) not used a chatbot before = 1	Daily = 0 4 – 6 times a week = 0 2 – 3 times a week = 0 Once a week = 4 Rarely = 13 Never = 0 Previously indicated (probably) not used a chatbot before = 13

*HBO = university of applied sciences
WO = academic university education

Figure C1.
Scree plot of the 42-item USIC for the complete participant group showing the Eigenvalue (variance) per factor

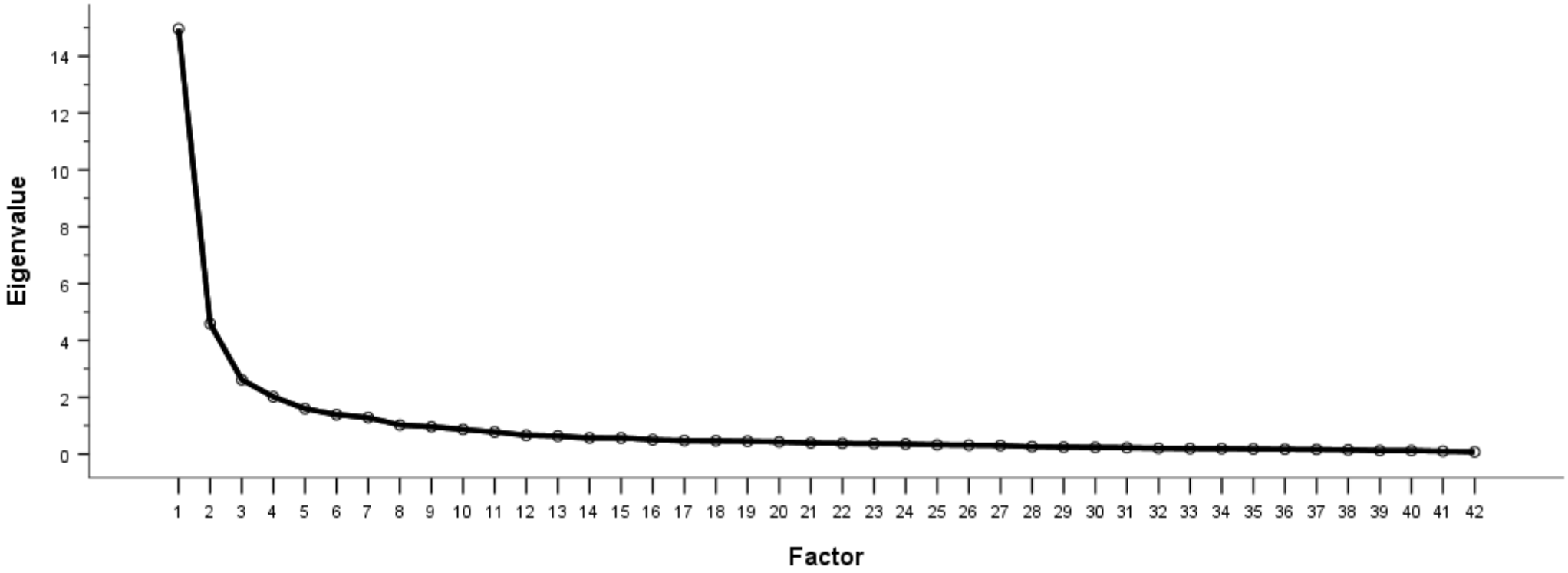


Table C2.
Correlation matrix of 42-item USIC

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
Q1	1.000																				
Q2	**0.664**	1.000																			
Q3	**0.601**	**0.650**	1.000																		
Q4	**0.613**	**0.616**	**0.541**	1.000																	
Q5	**0.648**	**0.649**	**0.548**	**0.757**	1.000																
Q6	**0.640**	**0.649**	**0.535**	**0.747**	**0.889**	1.000															
Q7	0.214	0.225	0.364	0.198	0.183	0.168	1.000														
Q8	0.159	0.123	0.244	0.100	0.150	0.128	0.488	1.000													
Q9	0.265	0.282	0.344	0.181	0.201	0.210	0.363	0.432	1.000												
Q10	0.014	0.009	0.086	-0.033	0.016	0.000	0.408	0.347	0.173	1.000											
Q11	-0.032	-0.053	0.071	-0.104	-0.038	-0.079	0.334	0.230	0.159	**0.602**	1.000										
Q12	0.084	0.097	0.236	0.015	0.084	0.086	**0.551**	0.380	0.356	**0.512**	0.444	1.000									
Q13	0.051	0.016	0.157	0.017	0.075	0.069	**0.551**	0.352	0.279	0.367	0.274	0.370	1.000								
Q14	0.120	0.149	0.237	0.083	0.062	0.069	**0.575**	0.485	0.364	0.434	0.329	0.460	0.495	1.000							
Q15	0.117	0.097	0.265	0.092	0.068	0.047	**0.580**	0.466	0.362	0.473	0.368	0.488	0.417	**0.674**	1.000						
Q16	0.102	0.121	0.257	0.060	0.029	0.037	**0.513**	0.379	0.299	0.373	0.323	0.442	0.294	**0.539**	0.677	1.000					
Q17	0.091	0.099	0.128	0.118	0.133	0.101	0.224	0.133	0.067	0.085	0.007	0.157	0.141	0.172	0.210	0.303	1.000				
Q18	0.156	0.210	0.328	0.154	0.143	0.159	0.473	0.373	0.353	0.344	0.246	0.420	0.280	**0.510**	0.534	**0.611**	0.383	1.000			
Q19	0.167	0.152	0.153	0.206	0.149	0.129	0.193	0.248	0.210	0.055	-0.004	0.124	0.216	0.302	0.210	0.145	0.145	0.268	1.000		
Q20	0.075	0.060	0.088	0.109	0.133	0.124	0.213	0.253	0.175	0.118	0.097	0.229	0.180	0.307	0.253	0.214	0.150	0.194	0.486	1.000	

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
Q21	0.114	0.086	0.154	0.124	0.096	0.071	0.192	0.263	0.183	0.070	0.011	0.146	0.198	0.288	0.205	0.129	0.180	0.210	0.789	0.486	1.000
Q22	0.039	0.041	0.175	-0.023	-0.030	-0.012	**0.567**	**0.503**	0.341	**0.530**	0.373	**0.543**	0.463	**0.655**	**0.751**	**0.614**	0.151	0.488	0.226	0.263	0.254
Q23	0.101	0.099	0.242	0.032	0.022	0.009	**0.569**	0.474	0.303	0.479	0.370	0.491	0.367	**0.629**	**0.750**	**0.778**	0.246	**0.574**	0.166	0.229	0.195
Q24	0.074	0.057	0.209	0.025	0.031	-0.004	**0.582**	**0.519**	0.301	0.518	0.398	0.531	0.450	**0.697**	**0.773**	**0.704**	0.214	**0.531**	0.225	0.303	0.265
Q25	0.066	0.063	0.178	-0.011	-0.006	0.034	**0.568**	0.446	0.344	0.476	0.378	0.476	0.470	**0.658**	**0.777**	**0.655**	0.154	**0.526**	0.173	0.271	0.191
Q26	0.052	0.040	0.184	0.027	-0.001	0.017	**0.598**	0.454	0.356	0.485	0.386	0.462	0.468	**0.646**	**0.766**	**0.640**	0.176	**0.529**	0.196	0.230	0.196
Q27	0.118	0.122	0.248	0.080	0.061	0.080	**0.612**	0.468	0.364	0.476	0.348	**0.506**	0.434	**0.641**	**0.778**	**0.756**	0.222	**0.585**	0.178	0.241	0.214
Q28	-0.040	0.040	0.086	-0.005	-0.053	-0.044	0.486	0.310	0.234	0.320	0.219	0.295	0.350	0.492	**0.572**	**0.522**	0.112	0.445	0.143	0.166	0.162
Q29	-0.005	0.031	0.132	-0.007	-0.033	-0.048	**0.504**	0.399	0.290	0.424	0.266	0.440	0.387	**0.561**	**0.652**	**0.552**	0.161	0.499	0.135	0.134	0.147
Q30	0.008	0.068	0.188	0.007	-0.019	-0.013	**0.507**	0.346	0.256	0.417	0.317	0.446	0.375	**0.542**	**0.651**	**0.518**	0.121	0.470	0.099	0.175	0.164
Q31	0.000	0.029	0.201	0.036	-0.009	0.006	**0.500**	0.301	0.265	0.322	0.317	0.365	0.399	**0.501**	**0.607**	**0.535**	0.110	0.479	0.173	0.265	0.211
Q32	-0.002	-0.032	0.088	-0.042	-0.062	-0.086	0.323	0.162	0.164	0.082	0.064	0.128	0.312	0.252	0.287	0.204	0.101	0.340	0.241	0.141	0.200
Q33	0.054	0.085	0.209	0.040	-0.016	0.009	0.376	0.216	0.253	0.131	0.133	0.198	0.255	0.395	0.467	0.374	0.097	0.416	0.250	0.135	0.255
Q34	0.120	0.191	0.259	0.115	0.090	0.069	**0.619**	0.392	0.332	0.351	0.261	0.379	0.419	**0.516**	**0.571**	0.438	0.181	0.462	0.220	0.128	0.226
Q35	0.145	0.199	0.299	0.138	0.120	0.130	**0.507**	0.293	0.347	0.282	0.233	0.347	0.404	0.498	**0.518**	0.357	0.200	**0.501**	0.230	0.135	0.191
Q36	0.182	0.208	0.202	0.120	0.091	0.075	0.472	0.287	0.374	0.197	0.177	0.300	0.304	0.432	0.433	0.298	0.077	0.416	0.168	0.065	0.146
Q37	0.113	0.220	0.257	0.135	0.063	0.071	0.464	0.391	0.423	0.271	0.239	0.296	0.253	**0.526**	**0.588**	0.484	0.250	0.494	0.243	0.198	0.244
Q38	0.143	0.193	0.239	0.160	0.099	0.087	0.296	0.186	0.297	0.048	0.091	0.180	0.238	0.403	0.362	0.276	0.242	0.338	0.367	0.265	0.386
Q39	0.139	0.252	0.276	0.173	0.111	0.103	0.506	0.371	0.353	0.238	0.232	0.308	0.251	**0.527**	**0.530**	0.467	0.249	**0.558**	0.333	0.231	0.322
Q40	0.102	0.049	0.123	0.101	0.021	0.035	0.132	0.103	0.152	-0.030	-0.063	0.131	0.073	0.074	0.120	0.062	0.082	0.129	0.155	0.040	0.124
Q41	0.123	0.186	0.274	0.120	0.071	0.085	0.256	0.255	0.238	0.027	-0.033	0.235	0.083	0.271	0.253	0.235	0.156	0.276	0.221	0.095	0.207
Q42	0.115	0.165	0.292	0.129	0.050	0.057	0.238	0.209	0.214	0.004	-0.026	0.197	0.094	0.224	0.225	0.223	0.162	0.272	0.167	0.123	0.157

	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	Q40	Q41	Q42
Q22	1.000																				
Q23	**0.754**	1.000																			
Q24	**0.835**	**0.812**	1.000																		
Q25	**0.739**	**0.677**	**0.741**	1.000																	
Q26	**0.712**	**0.714**	**0.742**	**0.789**	1.000																
Q27	**0.736**	**0.798**	**0.787**	**0.746**	**0.740**	1.000															
Q28	**0.567**	**0.583**	**0.580**	**0.570**	**0.609**	**0.640**	1.000														
Q29	**0.649**	**0.629**	**0.674**	**0.611**	**0.640**	**0.652**	**0.733**	1.000													
Q30	**0.610**	**0.558**	**0.614**	**0.627**	**0.642**	**0.635**	**0.580**	**0.717**	1.000												
Q31	**0.602**	**0.589**	**0.626**	**0.609**	**0.612**	**0.624**	0.496	**0.541**	**0.534**	1.000											
Q32	0.251	0.266	0.259	0.313	0.358	0.318	0.296	0.318	0.299	0.415	1.000										
Q33	0.348	0.383	0.409	0.404	0.459	0.468	0.425	0.394	0.416	**0.533**	**0.584**	1.000									
Q34	**0.538**	**0.500**	**0.544**	**0.509**	**0.549**	**0.550**	**0.541**	**0.616**	**0.581**	0.448	0.247	0.380	1.000								
Q35	0.429	0.419	0.432	0.449	**0.505**	0.493	0.472	**0.527**	**0.520**	0.395	0.299	0.357	**0.706**	1.000							
Q36	0.353	0.332	0.366	0.362	0.426	0.409	0.427	**0.515**	0.489	0.318	0.197	0.290	**0.642**	**0.732**	1.000						
Q37	**0.543**	**0.524**	**0.551**	**0.540**	**0.528**	**0.593**	0.434	**0.505**	0.589	0.494	0.243	0.359	**0.565**	0.479	0.410	1.000					
Q38	0.303	0.300	0.345	0.325	0.342	0.348	0.211	0.300	0.354	0.284	0.222	0.273	0.425	0.412	0.308	**0.572**	1.000				
Q39	0.473	0.478	0.469	**0.503**	0.498	**0.544**	0.402	**0.504**	0.516	0.416	0.223	0.364	**0.652**	**0.582**	**0.532**	**0.714**	**0.662**	1.000			
Q40	0.143	0.126	0.083	0.037	0.089	0.128	0.125	0.095	0.084	0.070	0.131	0.095	0.140	0.197	0.241	0.244	0.191	0.195	1.000		
Q41	0.213	0.238	0.231	0.204	0.226	0.259	0.183	0.252	0.165	0.181	0.172	0.192	0.279	0.327	0.418	0.264	0.225	0.321	**0.593**	1.000	
Q42	0.186	0.227	0.197	0.175	0.184	0.197	0.168	0.208	0.173	0.178	0.171	0.181	0.226	0.279	0.379	0.274	0.270	0.342	**0.627**	**0.844**	1.000

Table C3.
Correlation matrix of optimized 14-item USIC

	Q2	Q5	Q7	Q10	Q15	Q16	Q21	Q24	Q27	Q29	Q31	Q34	Q37	Q42
Q2	1.000													
Q5	**0.649**	1.000												
Q7	0.225	0.183	1.000											
Q10	0.009	0.016	**0.408**	1.000										
Q15	0.097	0.068	**0.580**	**0.473**	1.000									
Q16	0.121	0.029	**0.513**	0.373	**0.677**	1.000								
Q21	0.086	0.096	0.192	0.070	0.205	0.129	1.000							
Q24	0.057	0.031	**0.582**	**0.518**	**0.773**	**0.704**	0.265	1.000						
Q27	0.122	0.061	**0.612**	**0.476**	**0.778**	**0.756**	0.214	**0.787**	1.000					
Q29	0.031	-0.033	**0.504**	**0.424**	**0.652**	**0.552**	0.147	**0.674**	**0.652**	1.000				
Q31	0.029	-0.009	**0.500**	0.322	**0.607**	**0.535**	0.211	**0.626**	**0.624**	**0.541**	1.000			
Q34	0.191	0.090	**0.619**	0.351	**0.571**	**0.438**	0.226	**0.544**	**0.550**	**0.616**	**0.448**	1.000		
Q37	0.220	0.063	**0.464**	0.271	**0.588**	**0.484**	0.244	**0.551**	**0.593**	**0.505**	**0.494**	**0.565**	1.000	
Q42	0.165	0.050	0.238	0.004	0.225	0.223	0.157	0.197	0.197	0.208	0.178	0.226	0.274	1.000

Table C4.
Factor loadings for the principal component analysis of the 42-item USIC

		Factor			
Q#	Question	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q1	It was clear how to start a conversation with the chatbot.	**.812**	.033	.053	.062
Q2	It was easy for me to understand how to start the interaction with the chatbot.	**.821**	.049	.025	.144
Q3	I find it easy to start a conversation with the chatbot.	**.725**	.206	.055	.206
Q4	The chatbot was easy to access.	**.832**	-.025	.110	.067
Q5	The chatbot function was easily detectable.	**.899**	-.009	.066	-.052
Q6	It was easy to find the chatbot.	**.892**	-.010	.047	-.038
Q7	Communicating with the chatbot was clear.	.234	**.697**	.111	.185
Q8	I was immediately made aware of what information the chatbot can give me.	.169	**.522**	.214	.089
Q9	It is clear to me early on about what the chatbot can do.	.298	**.371**	.157	.232
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me. (R)	.019	**.664**	-.083	-.188
Q11	I had to pay special attention regarding my phrasing when communicating with the chatbot. (R)	-.041	**.556**	-.102	-.189
Q12	It was easy to tell the chatbot what I would like it to do.	.115	**.622**	.020	.054
Q13	The interaction with the chatbot felt like an ongoing conversation.	.046	**.541**	.187	.009
Q14	The chatbot was able to keep track of context.	.088	**.725**	.268	.132
Q15	The chatbot maintained relevant conversation.	.068	**.838**	.148	.141
Q16	The chatbot guided me to the relevant service.	.071	**.750**	.087	.097
Q17	The chatbot is using hyperlinks to guide me to my goal.	.126	.179	.234	.134
Q18	The chatbot was able to make references to the website o service when appropriate.	.181	**.606**	.199	.253
Q19	The interaction with the chatbot felt secure in terms of privacy.	.122	.072	**.854**	.106
Q20	I believe the chatbot informs me of any possible privacy issues.	.076	.189	**.675**	-.096
Q21	I believe that this chatbot maintains my privacy.	.056	.096	**.857**	.078
Q22	I felt that my intentions were understood by the chatbot.	-.026	**.832**	.171	.067
Q23	The chatbot was able to guide me to my goal.	.038	**.827**	.114	.095
Q24	I find that the chatbot understands what I want and helps me achieve my goal.	.011	**.858**	.195	.049
Q25	The chatbot gave relevant information during the whole conversation.	.002	**.834**	.151	.054
Q26	The chatbot is good at providing me with a helpful response at any point of the process.	-.006	**.836**	.144	.112
Q27	The chatbot provided relevant information as and when needed it.	.074	**.851**	.137	.132

Q28	The amount of received information was neither too much nor too less.	-.086	**.685**	.066	.202
Q29	The chatbot gives me the appropriate amount of information.	-.064	**.767**	.034	.244
Q30	The chatbot only gives me the information I need.	-.026	**.747**	.055	.197
Q31	The chatbot could handle situations in which the line of conversation was not clear.	-.034	**.685**	.211	.100
Q32	The chatbot explained gracefully when it could not help me.	-.130	.303	**.321**	.222
Q33	When the chatbot encountered a problem, it responded appropriately.	-.026	**.447**	.298	.224
Q34	I found the chatbot's responses clear.	.115	**.644**	.106	.367
Q35	The chatbot only states understandable answers.	.152	**.547**	.096	.457
Q36	The chatbot's responses were easy to understand.	.130	.455	-.013	**.568**
Q37	I feel like the chatbot's responses were accurate.	.108	**.585**	.246	.352
Q38	I believe that the chatbot only states reliable information	.119	.283	**.480**	.349
Q39	It appeared that the chatbot provided accurate and reliab information.	.150	**.528**	.314	.431
Q40	The time of the response was reasonable.	.018	-.021	.062	**.716**
Q41	My waiting time for a response from the chatbot was short.	.096	.119	.094	**.807**
Q42	The chatbot is quick to respond.	.084	.087	.083	**.821**

Note. Item's highest factor loading in boldface.

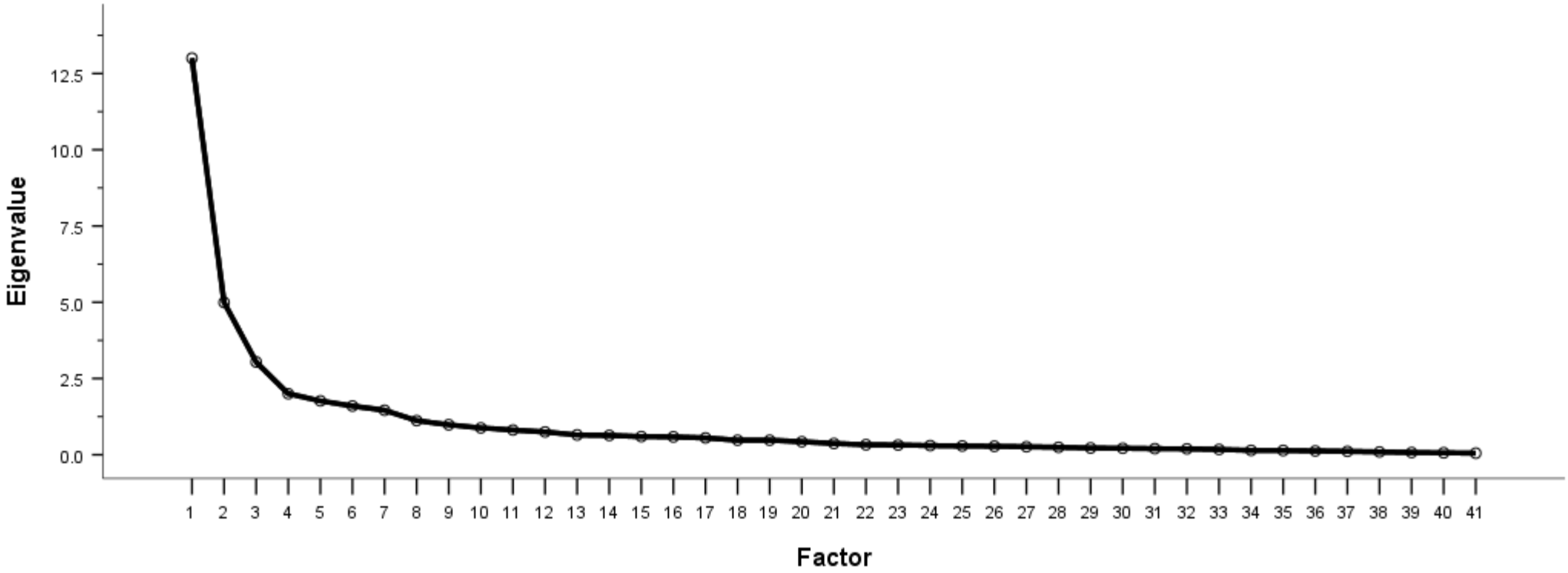
Table C5.
Factor loadings for the principal component analysis of the refined 33-item USIC with the associated features to identify the items with the highest factor loading per feature in a step towards the 14-item USIC

			Factor			
Q#	Question	Feature	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q1	It was clear how to start a conversation with the chatbot.	Ease of starting a conversation	**0.816**	0.045	0.055	0.031
Q2	It was easy for me to understand how to start the interaction with the chatbot.		**0.820**	0.059	0.006	0.163
Q3	I find it easy to start a conversation with the chatbot.		**0.727**	0.215	0.015	0.241
Q4	The chatbot was easy to access.	Accessibility	**0.842**	-0.002	0.093	0.052
Q5	The chatbot function was easily detectable.		**0.904**	0.001	0.057	-0.067
Q6	It was easy to find the chatbot.		**0.897**	0.003	0.029	-0.053
Q7	Communicating with the chatbot was clear.	Expectation setting	0.234	**0.709**	0.093	0.122
Q8	I was immediately made aware of what information the chatbot can give me.		0.145	**0.536**	0.246	0.079
Q9	I had to rephrase my input multiple times for the chatbot to be able to help me. (R)	Communication effort	0.002	**0.627**	-0.022	-0.213
Q12	It was easy to tell the chatbot what I would like it to do.		0.094	**0.613**	0.018	0.039
Q13	The interaction with the chatbot felt like an ongoing conversation.	Ability to maintain themed discussion	0.048	**0.557**	0.227	-0.116
Q14	The chatbot was able to keep track of context.		0.080	**0.747**	0.230	0.109
Q15	The chatbot maintained relevant conversation.		0.067	**0.858**	0.057	0.106
Q16	The chatbot guided me to the relevant service.	Reference to service	0.065	**0.763**	-0.052	0.133
Q18	The chatbot was able to make references to the website or service when appropriate.		0.182	**0.629**	0.101	0.256
Q19	The interaction with the chatbot felt secure in terms of privacy.	Perceived privacy	0.124	0.138	**0.906**	0.112
Q21	I believe that this chatbot maintains my privacy.		0.054	0.161	**0.902**	0.094
Q22	I felt that my intentions were understood by the chatbot.	Recognition and facilitation of user's goal and intent	-0.034	**0.854**	0.128	0.021
Q23	The chatbot was able to guide me to my goal.		0.035	**0.844**	0.012	0.090
Q24	I find that the chatbot understands what I want and helps me achieve my goal.		0.006	**0.878**	0.113	0.031
Q25	The chatbot gave relevant information during the whole conversation.	Relevance	0.000	**0.849**	0.054	0.035
Q26	The chatbot is good at providing me with a helpful response at any point of the process.		-0.004	**0.853**	0.063	0.061
Q27	The chatbot provided relevant information as and when I needed it.		0.076	**0.874**	0.030	0.096
Q28	The amount of received information was neither too much nor too less.	Maxim of quantity	-0.078	**0.708**	0.006	0.137
Q29	The chatbot gives me the appropriate amount of information.		-0.065	**0.785**	-0.013	0.182
Q30	The chatbot only gives me the information I need.		-0.021	**0.763**	-0.017	0.137

Q#	Question	Feature	Factor			
			Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q31	The chatbot could handle situations in which the line of conversation was not clear.	Graceful breakdown	-0.015	**<u>0.704</u>**	0.079	0.085
Q34	I found the chatbot's responses clear.	Understandability	0.109	**<u>0.664</u>**	0.131	0.285
Q35	The chatbot only states understandable answers.		0.151	**0.568**	0.116	0.359
Q37	I feel like the chatbot's responses were accurate.	Perceived credibility	0.103	**<u>0.625</u>**	0.151	0.322
Q39	It appeared that the chatbot provided accurate and reliable information.		0.144	**0.567**	0.251	0.424
Q41	My waiting time for a response from the chatbot was short.	Perceived speed	0.092	0.160	0.097	**0.857**
Q42	The chatbot is quick to respond.		0.084	0.130	0.044	**<u>0.876</u>**

Note. Item's highest factor loading in boldface and feature's highest factor loading underlined

Figure C2.
Scree plot of the 41-item USIC (excluding item Q17) for the 25-35 group showing the Eigenvalue (variance) per factor



Note. Item Q17 was removed during the assessment of the PCA's assumptions due to the lack of a moderate or strong correlation with other items

Figure C3.
Scree plot of the 42-item USIC for the 55-70 group showing the Eigenvalue (variance) per factor

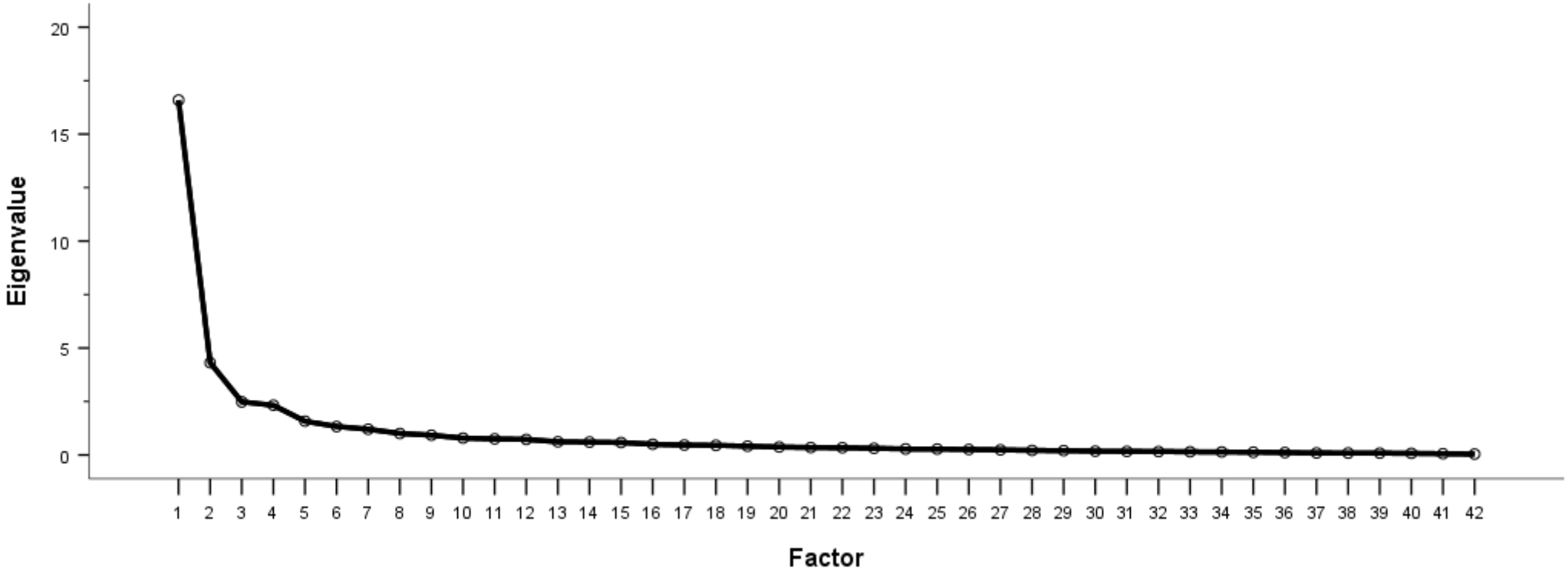


Table C6.
Factor loadings for the principal component analysis of the 41-item USIC (excluding item Q17) for participants between 25 and 35 of age

		Factor			
Q#	Question	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q1	It was clear how to start a conversation with the chatbot.	**0.816**	-0.001	0.090	-0.017
Q2	It was easy for me to understand how to start the interaction with the chatbot.	**0.821**	-0.078	0.153	-0.006
Q3	I find it easy to start a conversation with the chatbot.	**0.813**	0.085	0.095	0.063
Q4	The chatbot was easy to access.	**0.807**	-0.089	0.117	0.036
Q5	The chatbot function was easily detectable.	**0.895**	-0.040	-0.014	-0.093
Q6	It was easy to find the chatbot.	**0.899**	-0.037	-0.060	-0.086
Q7	Communicating with the chatbot was clear.	0.220	**0.628**	0.377	0.078
Q8	I was immediately made aware of what information the chatbot can give me.	0.233	**0.421**	0.186	0.052
Q9	It is clear to me early on about what the chatbot can do.	0.241	0.248	**0.402**	-0.050
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me. (R)	0.029	**0.774**	0.003	-0.116
Q11	I had to pay special attention regarding my phrasing when communicating with the chatbot. (R)	-0.093	**0.626**	0.052	-0.111
Q12	It was easy to tell the chatbot what I would like it to do.	0.064	**0.664**	0.136	0.029
Q13	The interaction with the chatbot felt like an ongoing conversation.	0.126	**0.494**	0.332	-0.092
Q14	The chatbot was able to keep track of context.	0.127	**0.641**	0.365	0.098
Q15	The chatbot maintained relevant conversation.	0.008	**0.774**	0.182	0.182
Q16	The chatbot guided me to the relevant service.	-0.010	**0.736**	0.052	0.268
Q18	The chatbot was able to make references to the website or service when appropriate.	0.100	**0.507**	0.091	0.322
Q19	The interaction with the chatbot felt secure in terms of privacy.	0.165	0.014	**0.496**	0.296
Q20	I believe the chatbot informs me of any possible privacy issues.	0.227	0.215	0.177	0.206
Q21	I believe that this chatbot maintains my privacy.	0.129	0.050	**0.510**	0.292
Q22	I felt that my intentions were understood by the chatbot.	-0.060	**0.846**	0.229	0.133
Q23	The chatbot was able to guide me to my goal.	-0.107	**0.812**	0.014	0.275
Q24	I find that the chatbot understands what I want and helps me achieve my goal.	-0.016	**0.877**	0.162	0.140
Q25	The chatbot gave relevant information during the whole conversation.	0.007	**0.809**	0.237	0.069
Q26	The chatbot is good at providing me with a helpful response at any point of the process.	-0.057	**0.787**	0.279	0.101
Q27	The chatbot provided relevant information as and when needed it.	0.018	**0.809**	0.149	0.185
Q28	The amount of received information was neither too much nor too less.	-0.201	**0.559**	0.343	0.081

		Factor			
Q#	Question	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q29	The chatbot gives me the appropriate amount of information.	-0.200	**0.632**	0.407	0.017
Q30	The chatbot only gives me the information I need.	-0.215	**0.558**	0.402	-0.002
Q31	The chatbot could handle situations in which the line of conversation was not clear.	-0.013	**0.608**	0.104	0.252
Q32	The chatbot explained gracefully when it could not help me.	-0.119	0.223	0.060	0.295
Q33	When the chatbot encountered a problem, it responded appropriately.	-0.113	**0.343**	0.091	0.292
Q34	I found the chatbot's responses clear.	0.031	0.415	**0.648**	-0.019
Q35	The chatbot only states understandable answers.	-0.031	0.246	**0.704**	0.016
Q36	The chatbot's responses were easy to understand.	-0.067	0.251	**0.659**	0.076
Q37	I feel like the chatbot's responses were accurate.	0.076	0.250	**0.636**	0.191
Q38	I believe that the chatbot only states reliable information	0.120	0.052	**0.680**	0.213
Q39	It appeared that the chatbot provided accurate and reliab information.	0.087	0.235	**0.739**	0.239
Q40	The time of the response was reasonable.	-0.098	0.065	0.264	**0.788**
Q41	My waiting time for a response from the chatbot was short.	0.066	0.123	0.168	**0.854**
Q42	The chatbot is quick to respond.	0.020	0.128	0.182	**0.852**

Note. Item's highest factor loading in boldface.

Table C7.
Factor loadings for the principal component analysis of the 42-item USIC for participants between 55 and 70 of age

		Factor			
Q#	Question	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q1	It was clear how to start a conversation with the chatbot.	**0.791**	0.054	0.040	0.101
Q2	It was easy for me to understand how to start the interaction with the chatbot.	**0.774**	0.141	-0.051	0.198
Q3	I find it easy to start a conversation with the chatbot.	**0.606**	0.321	-0.005	0.300
Q4	The chatbot was easy to access.	**0.855**	0.007	0.119	0.055
Q5	The chatbot function was easily detectable.	**0.898**	0.029	0.039	0.001
Q6	It was easy to find the chatbot.	**0.879**	0.038	0.046	0.044
Q7	Communicating with the chatbot was clear.	0.224	**0.723**	0.057	0.142
Q8	I was immediately made aware of what information the chatbot can give me.	0.113	**0.597**	0.214	0.107
Q9	It is clear to me early on about what the chatbot can do.	0.295	0.435	0.103	0.305
Q10	I had to rephrase my input multiple times for the chatbot to be able to help me. (R)	0.057	**0.577**	-0.186	-0.210
Q11	I had to pay special attention regarding my phrasing when communicating with the chatbot. (R)	0.010	0.468	-0.148	-0.360
Q12	It was easy to tell the chatbot what I would like it to do.	0.191	**0.589**	0.009	0.079
Q13	The interaction with the chatbot felt like an ongoing conversation.	0.007	**0.558**	0.190	0.013
Q14	The chatbot was able to keep track of context.	0.031	**0.773**	0.246	0.094
Q15	The chatbot maintained relevant conversation.	0.094	**0.893**	0.129	0.066
Q16	The chatbot guided me to the relevant service.	0.101	**0.785**	-0.009	0.015
Q17	The chatbot is using hyperlinks to guide me to my goal.	0.112	0.275	0.070	0.255
Q18	The chatbot was able to make references to the website o service when appropriate.	0.163	**0.731**	0.080	0.209
Q19	The interaction with the chatbot felt secure in terms of privacy.	0.105	0.108	**0.860**	0.081
Q20	I believe the chatbot informs me of any possible privacy issues.	0.081	0.201	**0.764**	-0.086
Q21	I believe that this chatbot maintains my privacy.	0.038	0.124	**0.832**	0.046
Q22	I felt that my intentions were understood by the chatbot.	0.029	**0.806**	0.199	-0.001
Q23	The chatbot was able to guide me to my goal.	0.134	**0.859**	0.089	-0.020
Q24	I find that the chatbot understands what I want and helps me achieve my goal.	0.046	**0.843**	0.249	-0.019
Q25	The chatbot gave relevant information during the whole conversation.	0.036	**0.858**	0.164	0.031

		Factor			
Q#	Question	Conversation start 1	Communication quality 2	Perceived privacy 3	Perceived speed 4
Q26	The chatbot is good at providing me with a helpful response at any point of the process.	0.032	**0.853**	0.159	0.039
Q27	The chatbot provided relevant information as and when needed it.	0.086	**0.893**	0.136	0.053
Q28	The amount of received information was neither too much nor too less.	-0.014	**0.746**	0.083	0.070
Q29	The chatbot gives me the appropriate amount of information.	-0.002	**0.819**	0.042	0.157
Q30	The chatbot only gives me the information I need.	0.077	**0.829**	0.052	0.110
Q31	The chatbot could handle situations in which the line of conversation was not clear.	-0.017	**0.760**	0.271	-0.032
Q32	The chatbot explained gracefully when it could not help me.	-0.124	0.406	0.400	0.190
Q33	When the chatbot encountered a problem, it responded appropriately.	0.022	**0.581**	0.337	0.136
Q34	I found the chatbot's responses clear.	0.046	**0.759**	0.008	0.282
Q35	The chatbot only states understandable answers.	0.110	**0.697**	0.007	0.363
Q36	The chatbot's responses were easy to understand.	0.139	**0.543**	-0.068	0.517
Q37	I feel like the chatbot's responses were accurate.	0.025	**0.761**	0.102	0.233
Q38	I believe that the chatbot only states reliable information	0.028	0.416	0.411	0.279
Q39	It appeared that the chatbot provided accurate and reliab information.	0.075	**0.684**	0.154	0.322
Q40	The time of the response was reasonable.	0.103	-0.059	0.010	**0.657**
Q41	My waiting time for a response from the chatbot was short.	0.118	0.186	0.047	**0.842**
Q42	The chatbot is quick to respond.	0.175	0.117	0.071	**0.859**

Note. Item's highest factor loading in boldface.