Assessing User Satisfaction with Information Chatbots: A Preliminary Investigation

Divyaa Balaji

Submitted in Partial Fulfillment of Requirements for Master of Science in Psychology
(Human Factors and Engineering Psychology)

University of Twente

Abstract

Despite the increasing number of service chatbots available today, it has been observed that many often fail to impress their customers (Brandtzaeg & Folstad, 2018). In order to provide a better experience for end-users of information chatbots such as those designed for customer service, chatbot developers can benefit from a diagnostic measure of user satisfaction. This thesis follows the work conducted by Tariverdiyeva and Borsci (2019) towards the development of a diagnostic questionnaire that provides an assessment of user satisfaction with information chatbots. A pre-experimental phase was undertaken in which the original list of chatbot features obtained from Tariverdiyeva and Borsci (2019) was reviewed by a team of experts and an extended literature review was conducted to ensure that all relevant chatbot features had been identified. The resulting list of chatbot features was used to generate an item pool. Study 1 reports the results of a series of focus groups in which participants discussed the updated list of chatbot features and the corresponding item pool based on which further refinement took place. Study 2 describes steps taken towards a preliminary evaluation of the questionnaire. The item pool was administered to a sample of 60 university students and analyses were conducted in order to test the questionnaire's underlying factor structure and reliability. It was found that the data acquired from participants can be captured by four factors - communication quality, interaction quality, perceived privacy and perceived speed. Actions for future studies are discussed in order to arrive at the desired questionnaire.

Keywords: chatbot, conversational interface, user satisfaction, usability, user expectations

Acknowledgements

I would like to express my gratitude to,

Dr. Simone Borsci, whose door was always open if I needed guidance. Thank you for being supportive throughout the whole process, approachable at any time of the day and willing to openly listen to my ideas and doubts and provide feedback;

Dr. Martin Schmettow, for taking the time to help me understand advanced concepts and providing critical feedback that allowed me to refine my analysis;

And to Nina Bocker and Lisa Waldera for helping me gather participants and collect not only enough data for their own theses but also enough for me to reach my own target.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1. Introduction

1.1. The rise of chatbots

Chatbots are a class of intelligent conversational web- or mobile software applications that can engage in a dialogue with humans using natural language (Radziwill & Benton, 2017). Unlike voice assistants, which allow users to complete a range of actions simply by speaking commands, chatbots more commonly rely on text-based interactions and are primarily being implemented as service-oriented bots by businesses on websites and other instant messaging platforms to answer customer queries and help them navigate their service. This is consistent with information-type chatbots (Paikari & van der Hoek, 2018) which serve the purpose of helping the user find information that may be relevant to the task at hand. Commonly cited benefits for the adoption of chatbots by businesses include reduced operational costs associated with customer service, the opportunity to cultivate an effective brand image and the potential to reach a staggering number of customers with ease. In fact, chatbots have already gained considerable traction among users. 67% of customers worldwide have used a chatbot for customer support in 2017 (LivePerson, 2017) and approximately 40% of millennials chat with chatbots daily (Acquire.io, 2018). Furthermore, it is predicted that by 2020, 85% of customer interactions will be handled without a human agent (Chatbots Life, 2019).

Chatbots have been around since the 1960's but have only recently gained the attention of businesses and their consumers, and this is likely because of two main trends. First, the progress that has been made in the field of artificial intelligence has given rise to the technology that allows chatbots to understand and respond intelligently to an impressive range of natural language input. Secondly, the changes that have taken place in the way we communicate today have created an environment in which conversational interfaces such as chatbots can truly flourish. Specifically, people all over the world of all ages are significantly more comfortable

communicating through the short-typed interactions characteristic of instant messaging. Men and women between the ages of 18 to 44 years comprise of approximately 75% of Facebook users worldwide and Facebook reported that 2.7 billion people were using at least one of the company's core products (Facebook, WhatsApp, Instagram or Messenger) every month (Statista, 2019). Consequently, potential end-users of chatbots would likely learn how to use chatbots very quickly, having already been accustomed to this manner of conveying and receiving information. Consistent with this notion, it has been suggested by others that chatbots possess superior flexibility and ease of use compared to web- and mobile-based applications and could soon replace them to become the *universal user interface* (Solomon, 2017).

1.2. The need for a measure of user satisfaction with chatbots

As service bots are becoming more commonplace, there is a growing need to assess user satisfaction with these applications as the hypothesised benefits of adopting information chatbots in lieu of human customer service can only be realised if customers are willing to continuously engage with such bots and experience these interactions positively. It is therefore unsurprising that recent studies involving chatbots have assessed user satisfaction in one way or another. For example, a study by Morris, Kouddous, Kshirsagar & Shueller (2018) asked participants to simply rate each response on a single-item, three-point Likert scale (good, ok, bad) and a similar approach was utilised by Skjuve et al. (2019) in which overall perceived *pleasantness* was assessed with a single open-ended free-text follow-up question. A 15-item questionnaire comprising questions related to seven subjective metrics (usage, task ease, interaction pace, user expertise, system response, expected behaviour and future use) was utilised in two other studies (Walker, Passonneau & Boland, 2001; Steinbauer, Kern & Kroll, 2019). Alternatively, Trivedi (2019) looked at chatbots as a type of information system and thus used a measure of information system user satisfaction (Brown & Jayakody, 2008). There are just a few examples that show the significant variability in the types of questions posed to

participants, almost all of which have been devised by the individual researcher in the apparent absence of a standardised approach.

There are several standardised measures of perceived usability such as SUS (Brooke, 1996), UMUX (Bosley, 2013), UMUX-LITE (Lewis, Utesch & Maher, 2013) and CSUQ (Lewis, 2002) and have been shown to be valid and reliable across different contexts and interfaces (Lewis, 2018; Borsci, Federici, Bacci, Gnaldi & Bartolucci, 2015). However, the observation that many researchers have resorted to devising their own questionnaires instead of utilising these existing measures suggests a hidden assumption in the field that these measures may not be appropriate in the context of chatbots. One reason for this may be the fact that these measures of perceived usability are non-diagnostic – while they can indicate overall usability of the system, they cannot provide information on specific aspects of the system. As chatbot technology is in the infancy of its adoption life cycle, diagnosticity may prove to be an important requirement for current user satisfaction assessment, playing a critical role in informing designers about the specific aspects of the chatbot interaction that can be improved in order to provide a better user experience for customers.

Another explanation is provided by Folstad and Brandtzaeg (2018), who point out that natural-language user interfaces chatbots depart significantly from traditional interfaces in that the object of design is now the conversation itself. With graphical user interfaces, designers had a large degree of control over how the content and features are presented to the user in the process of designing the visual layout and interaction mechanisms. On the other hand, a natural-language interface is largely a “blank canvas” - the underlying features and content are hidden from the user and the interaction therefore critically hinges on user input. The key success factor for natural-language interfaces lies in their ability to “support user needs in the conversational process seamlessly and efficiently” (Brandtzaeg & Folstad, 2017). Given the unique challenge posed by natural-language interfaces, it is highly likely that the factors that

contribute to user satisfaction with information-retrieval chatbots are different thus requiring a different approach to assessment (Piccolo, Mensio & Alani, 2019).

Tariverdiyeva and Borsci (2019) provide additional support for the relative inadequacy of existing measures. The authors compared the usability of websites with their chatbot counterparts by administering the 2-item UMUX-lite (Lewis et al., 2013) after instructing participants to perform the same information-retrieval task with both interfaces. It was concluded that while existing measures such as the UMUX-lite can be a good indicator of overall usability, a tool that provides more diagnostic information about the interaction with the chatbot by assessing additional aspects of the interaction would benefit the designer's understanding and decision making. In conclusion, there is a need for a valid, reliable measurement tool to assess user satisfaction with text-based information chatbots that can be utilised by both business and researchers to evaluate interaction quality in a short yet informative manner.

1.3. Previous work

Tariverdiyeva and Borsci (2019) initiated work in this area by conducting a qualitative systematic literature review to explore the features that could influence users' perceptions of chatbots. The review yielded 27 different features that could be relevant in informing user satisfaction with chatbots and other conversational agents. These features were then presented in an online survey directed at end-users and experts, who were asked to provide their opinions on how important they considered each feature to be in the context of chatbot interactions. Upon computing consensus across groups for each feature and considering other comments made by users, the list was reduced to 18 features (Table 1). Those marked with an asterisk (*) were found to be the most important chatbot features based on full consensus across all groups.

Several limitations regarding the study were noted. Firstly, there was a significant difference between experts and end-users in the relative importance assigned to different features. As the construct in question is that of user satisfaction, it may be pertinent to further validate the findings of this study with an emphasis on the opinions of potential end-users. Additionally, it was acknowledged that it could not be known that all respondents interpreted the features and their descriptions as was intended which could have skewed the results. Given that the literature review was based on a specific set of keywords and sample sizes utilized in the study were small, it is possible that several factors relevant to the construct of user satisfaction were overlooked, resulting in inadequate content validity.

T**able 1:** *List of 18 chatbot features obtained from Tariverdiyeva and Borsci (2019)*

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

1.4. Present study

The present study aims to address the above limitations and build on previous work (Tariverdiyeva & Borsci, 2019) by developing a diagnostic questionnaire to assess user satisfaction with information chatbots (USIC) in three phases:

i. The pre-experimental phase will corroborate and build upon previous findings. This phase will consist of three activities. First, a research team comprising three experts will review the list of 18 features that Tariverdiyeva and Borsci (2019) arrived at. Secondly, an extended literature review will be carried out using a different set of search terms to identify relevant features that may have been overlooked in the previous study. This will result in a preliminary revised list of features. Thirdly, once the research team reaches a consensus on the content adequacy of the revised list of features, questionnaire items will be generated for each of these features to generate a preliminary item pool.

ii. Study 1 will involve a series of focus groups will be conducted using potential end-users of chatbots in order to (a) obtain an in-depth understanding of the features are important (or not) in determining their satisfaction with information-retrieval chatbots in order to confirm that the preliminary list of revised features captures the construct adequately and (b) obtain feedback on the item pool. The list of features and item pool will be reviewed based on data gathered from the focus groups.

iii. Study 2 will then execute usability tests with different chatbots during which the preliminary item pool will be administered to potential end-users as a post-test questionnaire. Analyses will be used to uncover the underlying factor structure to provide preliminary evidence to support the questionnaire's validity and reliability.

2. Pre-experimental Phase

2.1. Review of initial list of features

The initial list of features obtained by Tariverdiyeva and Borsci (2019) was qualitatively reviewed by a research team comprising of three experts. Each feature was discussed along the following questions - (a) what exactly does this feature refer to in an interaction with an information chatbot? (b) how and why would it be important in determining user satisfaction? and (c) thus, is it truly relevant to user satisfaction with information chatbots?

Trust and *ease of use* were two features that resulted in significant discussion among the research team. Upon exploring what each of these features meant in the context of a chatbot interaction, it was quickly discovered that both features are likely multidimensional and thus too broad to be captured by single features. These two features were re-conceptualized and separated into more specific component features. However, the initial broad features were also retained in addition to the component features described below as the research team wanted to confirm through the subsequent series of focus groups whether making such distinctions is a valid approach to user satisfaction.

For example, *trust* can apply to different aspects of a chatbot interaction. One expert proposed that users must feel like they can trust the chatbot, particularly information retrieval chatbots, to provide them with accurate and reliable information (Luger & Sellen, 2016). Another expert offered that users must also feel like they can trust the chatbot to safeguard their privacy and handle personal data securely. This notion is consistent with an exploratory study that found that trust in chatbots was informed not only by the quality with which it interpreted users' requests and the advice it provided but also the perceived security and privacy associated with the service context (Folstad, Nordheim & Bjokli, 2018). More importantly, it was agreed that these two aspects of trust are likely independent, making it important that such a distinction

be made. *Trust* was replaced with two new features that captured the two aspects of trust that arose during discussion, namely *perceived credibility* and *privacy & security*. It was noticed that *perceived credibility* was similar to *maxim of quality* which is included in the initial list and refers to the accuracy of information that is provided to the user. When these two features were reviewed, it was agreed that the user has no way of knowing whether the information given is accurate or not (*maxim of quality*) but can still form a subjective opinion about the same (*perceived credibility*) and it seemed more likely that this perception would significantly determine end-user satisfaction independent of the information's genuine accuracy. *Maxim of quality* was thus excluded and replaced by *perceived credibility*.

A similar discussion arose for *ease of use* when the research team explored what it means for a chatbot to be easy to use. Members began by listing the various ways in which a chatbot could be considered easy to use. As the discussion progressed, it became apparent that ease of use could mean different things in the course of an entire interaction with a chatbot from start to finish (Zamora, 2017), suggesting that it may be worthwhile to explore the possibility that ease of use may be composed of different, more specific features. For example, the user should find it easy to find the chatbot (*visibility*) as well as start a conversation with it (*ease of starting a conversation*). Users may also expect to be able to easily convey their wishes to the chatbot in however they choose to phrase their input and importantly, avoid putting in too much effort and rephrasing so that the chatbot may understand (*flexibility of linguistic input*). Additionally, the output produced by the chatbot must be clear and easy to interpret for the user (*understandability*; renamed from *maxim of manners*). *Maxim of manners* was reconceptualised as *understandability* because the original definition for *maxim of manners* addressed not only the clarity of the response but also its conciseness, which appears to have already been addressed by *maxim of quantity* in that the information presented must be of the appropriate amount.

Additionally, it was agreed to reconceptualise and rename two features so that they reflected the intended chatbot feature more accurately. Firstly, it was felt that *appropriate degrees of formality* only addressed one aspect of a much larger concept, that is, the way in which the chatbot uses language to communicate. Chatbots may additionally also employ the right vocabulary, tone and other general mannerisms, contributing to its language style as a whole. As language style had not been captured by any of the other existing features in the initial list, this feature was renamed as *appropriate language style* to encompass all the above aspects. *Reference to what is on the screen* emerged as a somewhat confusing feature to the research team as it was pointed out that chatbots exist on multiple platforms including instant messaging platforms like Facebook and WhatsApp. While this feature may be relevant for chatbots embedded on websites, it is not always possible for a chatbot to make a reference to something that is on the screen. However, the team agreed that making references to the business it serves is indeed important. While these references could be directed at the screen itself, it can also include hyperlinks provided as part of the response as well as automatic transitions to certain webpages. Based on the discussion, the feature was renamed to *reference to service* and thus includes any kind of reference that the chatbot makes to the service it operates for. However, as *reference to service* includes references made within and to webpages, this made the feature very similar to *integration with the website*. However, *reference to service* not only covers the extent to which the chatbot is integrated with the website, it includes other forms of reference too therefore it was agreed to subsume *integration with the website* under *reference to service*.

Finally, two features were excluded from the initial list: *ethical decision-making* and *meeting of neuro-diverse needs*. Initially, experts agreed that if asked, users would indeed expect chatbots to exhibit the above characteristics, making these features apparently relevant to assessing end-user satisfaction with chatbot interactions. However, the measurement tool in

development is being targeted at the single user and the experts quickly realized that a single user would not be able to evaluate a given information-retrieval chatbot along these two features based on his or her interaction alone. For example, not every chatbot interaction would warrant an ethical decision to be made and similarly, whether the chatbot meets neuro-diverse needs would also be difficult for a single user to evaluate after his or her interaction. Upon discussion, the experts concluded that it would be difficult for a user to evaluate a given information-retrieval chatbot along these two factors. The experts further agreed that while the above features might not be relevant for evaluating end-user satisfaction, they remain relevant for chatbot design and could thus inform a checklist directed at designers consisting of features that every chatbot should incorporate for success across different user groups.

Table 2 summarizes the changes made to the original list of chatbot features (Table 1) and presents an updated list of chatbot features with their descriptions. Chatbot features that were modified in some way are clarified under *refined chatbot feature* - chatbot features that remain unchanged have no counterpart under this column. Chatbot features that were removed from the original list are marked by '(R)' beside the relevant original feature.

Table 2*: List of chatbot features after review by research team*

Original chatbot feature	Refined chatbot feature	Description
Response time		Ability of the chatbot to respond timely to users' requests
Graceful responses in unexpected situations		Ability of the chatbot to gracefully handle unexpected input, communication mismatch and broken line of conversation
Maxim of quantity		Ability of the chatbot to respond in an informative way without adding too much information
Recognition and facilitation of users' goal and intent		Ability of the chatbot to understand the goal and intention of the user and to help them accomplish these
Maxim of quality (R)	*Refer to*: perceived credibility	
Perceived ease of use (R)	Ease of use (general)	How easy the user feels it is to interact with the chatbot
	Visibility	How easy it is to locate and spot the chatbot
	Ease of starting a conversation	How easy the user feels it is to start interacting with the chatbot and start typing
Maxim of manners	Understandability	Ability of the chatbot to communicate clearly in such a way that it is easily understandable
Engage in on-the-fly problem solving		Ability of the chatbot to solve problems instantly on the spot
Maxim of relation		Ability of the chatbot to provide relevant and appropriate contributions to users' needs at each stage
Themed discussion		Ability of the chatbot to maintain a conversational theme once introduced and keep track of context to understand user input
Appropriate degrees of formality	Appropriate language style	Ability of the chatbot to use the appropriate language style for the context
Users' privacy and ethical decision making (R)	*Refer to*: privacy & security	
Reference to what is on the screen Integration with the website	Reference to service	Ability of the chatbot to make references to the relevant service, for example, by providing links or automatically navigating to pages
Meets neurodiversity needs (R)		
Trustworthiness (R)	Trust (general)	Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage
	Perceived credibility	How correct and reliable the chatbot's response seems to be
	Privacy & security	The extent to which the user feels that the interaction with the chatbot is secure and protects their privacy
Process facilitation and follow up	Process tracking	Ability of the chatbot to inform and update users about the status of their task in progress
Flexibility of linguistic input		How easily the chatbot understands the user's input

2.2. Extended literature review

2.2.1. Introduction.

The systematic literature review conducted by Tariverdiyeva and Borsci (2019) focused on studies that included theories or experimental findings on factors that were potentially relevant in determining user satisfaction and perceived usability with information chatbots. Subsequently, the search terms used were: “conversational interface”, “conversational agent”, “chatbot”, “interaction”, “quality”, “satisfaction”. In light of the authors’ acknowledgment that this list may not be complete, the extended literature review served two objectives: (a) to identify chatbot features that are not present in the list of 18 chatbot features obtained from Table 1 and (b) to do so by using a different set of search terms that instead focused on studies which investigated end-user needs, expectations and motivations in the context of chatbots given the need for a more user-centred approach to chatbot interaction assessment and design.

2.2.2. Method.

The systematic literature review was qualitative and followed the method put forth by Ogawa and Malen (1991). The search was conducted through Google Scholar using the following search string: “chatbots” “user” “expectations”. Given the explosion of chatbot-related studies in the last few years (Piccolo, Mensio & Alani, 2019), the search was limited to articles within the last five years. The search yielded a total of 1,810 results. Inclusion criteria for screening based on abstract was focused on articles that (a) explicitly explored or identified, in some way, end-user expectations for different chatbots with a focus on customer-service/information-retrieval chatbots and (b) addressed features of chatbots that were not present in Table 1.

As the number of articles to screen was too large, the principle of inductive thematic saturation (Saunders et al., 2018) was used to limit the number of articles screened such that

sampling of articles was halted upon discovering that additional articles did not provide indications of new chatbot features that had not already been found. In this review, pages of results were scanned one at a time and the articles on each page were screened on the basis of abstract to determine the article's relevance to the review. As the review progressed, the number of relevant articles found on a given page was zero and remained as such for consecutive pages of results, showing evidence of saturation. (Figure 1). Additionally, we became convinced that the articles screened thus far have satisfactorily served the review purpose and captured any additional chatbot features that may have been excluded in the prior literature review. Thus, based on saturation, it was deemed that the sampling of articles could be halted, and the review can proceed systematically with the number of articles screened hitherto (n = 260). Full-text articles of the articles shortlisted based on abstract (n = 38) were examined for their usefulness to the review, yielding 23 articles that were utilised in the qualitative synthesis. A flow diagram of the review process is depicted in Figure 2.

Figure 1*: Evidence of saturation found during sampling of articles for screening during extended literature review*

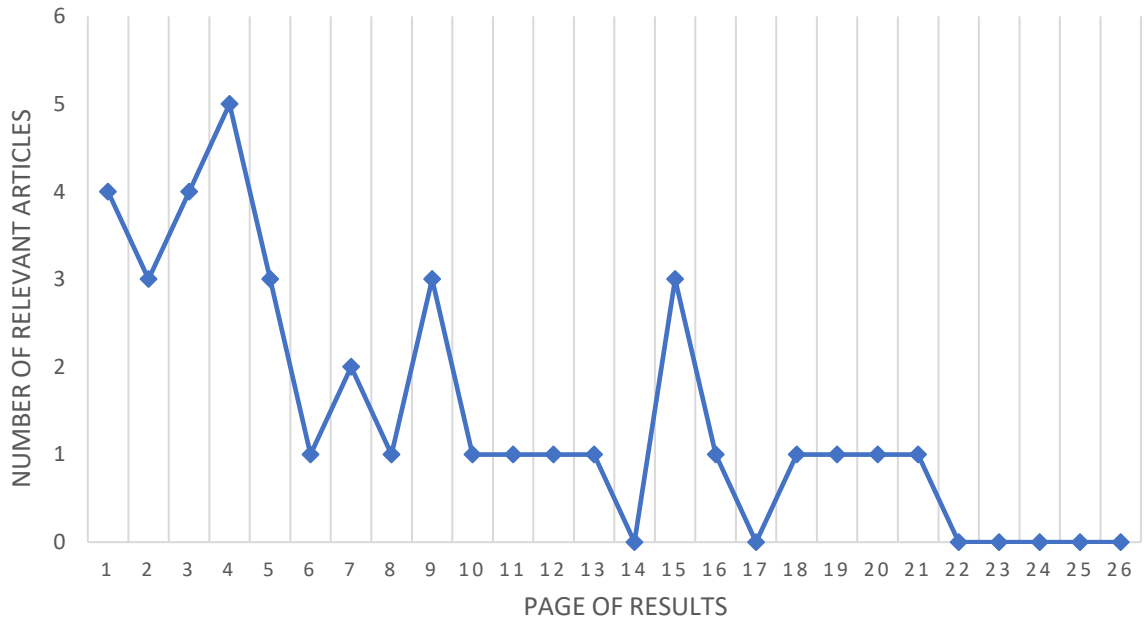
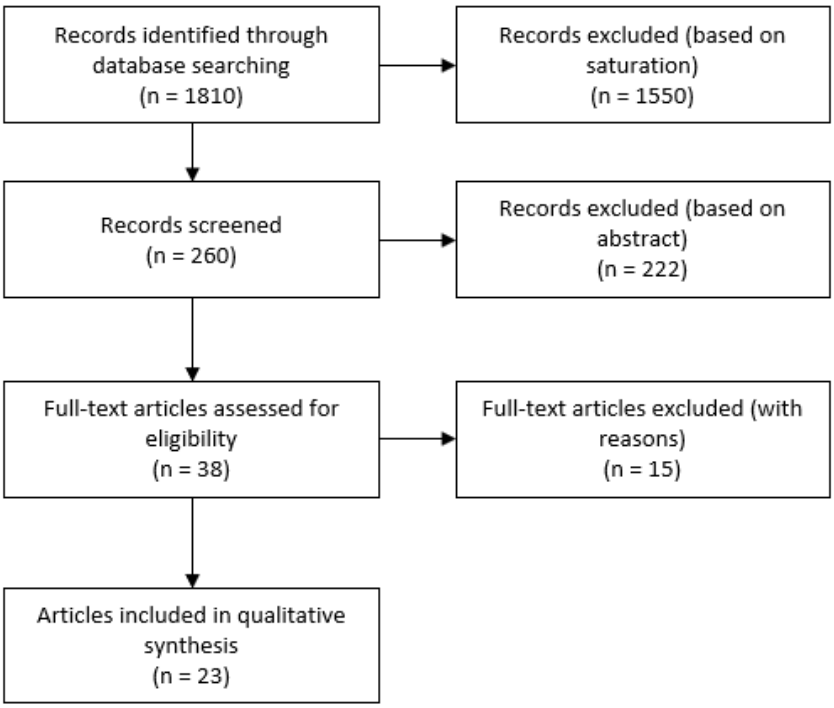


Figure 2*: PRISMA flow diagram depicting systematic review*



2.2.3. Results.

A qualitative synthesis of the selected articles revealed three additional chatbot features that were not present in Table 1 and may play an important role in shaping user satisfaction with information chatbots. A summary of the chatbot features and the relevant articles can be found in Table 3. Additionally, short rationales for the inclusion of each of these chatbot features based on the relevant literature are presented below.

Expectation setting. It has been found that a feature of successful chatbots is expectation setting, or the act of informing users about what to expect from the subsequent interaction. This includes not only full transparency about the fact that the user is interacting with a chatbot and not a human and what the chatbot can and cannot deliver. If this information is not provided upfront, users tended to either overestimate or underestimate the chatbot's capabilities which, often, results in user confusion and frustration. Chatbot actions that indicate the extent of the chatbot's capabilities tend to set realistic expectations increasing user comfort and ease of use.

Personality. A significant body of literature documents the benefit of paying attention to the chatbot's personality in pursuit of a positive user experience. This may largely be since as a type of natural language interface, one of the drivers of chatbot use is the possibility of interacting with the bot as one would with another human-being. In addition to providing a more natural conversational experience, the inclusion of a personality can significantly contribute to the *human-likeness* of the chatbot, which has been shown to play an important role in the degree to which the user accepts and trusts the chatbot. This feature was, however, found to be unimportant and was eliminated in the study conducted by Tariverdiyeva and Borsci (2019). Upon discovering that personality repeatedly emerged as an important determinant of user satisfaction across numerous studies in the present extended literature

review, it was decided to include this feature with the intent of confirming the (un)importance of this feature with certainty.

Enjoyment. Another feature that emerged as relevant to user experience with chatbots is the extent to which the user is engaged with and enjoys the interaction. Promoting fun and playful experiences involving humour among other diverse responses has been posited as a desirable characteristic for interfaces in order to promote adoption and satisfaction. Playful interactions, especially at the initial stages of interaction, are often thought to be engaging as *points of entry* to the system which encouraged sustained use of the chatbot and allowed users to be more forgiving of failures in the earlier stages.

Table 3*: Additional chatbot features obtained from extended literature review*

Chatbot feature	Description	References
Expectation Setting	The extent to which the chatbot sets expectations for the interaction with an emphasis on what it can and cannot do	Brandtzaeg & Folstad (2018a); Jain, Kumar, Kota & Patel, 2018; Luger & Sellen (2016); Chopra & Chivukula (2017); Sorensen (2017); Go & Sundar (2019)
Personality	The chatbot presents to the user a pleasant and human-like personality during the interaction	Jain, Kumar, Kota & Patel (2018), Zamora (2017), Chopra & Chivukula (2017); Peras (2018); Lannoy (2017); de Haan et al. (2018); Diederich et al. (2019); Piccolo, Mensio & Alani (2019), Assink (2019), Sheehan (2018); Smestad & Volden (2018); Hendriks (2019); Yang, Aurisicchio & Baxter (2019); Folstad & Skjuve (2018), Verney & Poulain (2018)
Enjoyment	The extent to which the user has an enjoyable and engaging interaction with the chatbot	Liao et al. (2018); Jain, Kumar, Kota & Patel (2018); Luger & Sellen (2016); Muresan & Pohl (2019); Piccolo, Mensio & Alani (2019); Yang, Aurisicchio & Baxter (2019); Nijholt, Niculescu, Alessandro & Banchs (2017); Folstad & Skjuve (2018)

Table 4*: Revised list of 21 chatbot features at the end of pre-experimental phase review (in no particular order)*

	Chatbot feature	Description
1	Response time	Ability of the chatbot to respond timely to users' requests
2	Graceful responses in unexpected situations	Ability of the chatbots to gracefully handle unexpected input, communication mismatch and broken line of conversation
3	Maxim of quantity	Ability of the chatbot to respond in an informative way without adding too much information
4	Recognition and facilitation of users' goal and intent	Ability of the chatbot to understand the goal and intention of the user and to help him accomplish these
5	Perceived credibility	How correct and reliable the chatbot's output seems to be
6	Ease of use (general)	How easy the user feels it is to interact with the chatbot
7	Engage in on-the-fly problem solving	Ability of the chatbot to solve problems instantly on the spot
8	Maxim of relation	Ability of the chatbot to provide the relevant and appropriate contribution to people’s needs at each stage
9	Ability to maintain themed discussion	Ability of the chatbot to maintain a conversational theme once introduced and keep track of context to understand user input
10	Appropriate language style	Ability of the chatbot to use appropriate language style for the context
11	Reference to service	Ability of the chatbot to make references to the relevant service, for example, by providing links or automatically navigating to pages.
12	Trust (general)	Ability of the chatbot to convey accountability and trustworthiness to increase willingness to engage
13	Process tracking	Ability of the chatbot to inform and update users about the status of their task in progress
14	Flexibility of linguistic input	How easily the chatbot understands the user's input
15	Privacy & security	The extent to which the user feels that the interaction with the chatbot is secure and protects their privacy
16	Understandability	Ability of the chatbot to communicate clearly and is easily understandable
17	Visibility	How easy it is to locate and spot the chatbot
18	Ease of starting a conversation	How easy it is to start interacting with the chatbot and start typing
19	Expectation setting	The extent to which the chatbot sets expectations for the interaction with an emphasis on what it can and cannot do
20	Personality	The chatbot presents to the user a pleasant and human-like personality during the interaction
21	Enjoyment	The extent to which the user has an enjoyable and engaging interaction with the chatbot

2.3. Generation of item pool

Table 4 shows the list of 21 chatbot features arrived at after reviewing the initial list of features (Table 2) and identifying additional chatbot features (Table 3). At the time of item generation, the underlying factor structure was unknown and thus it was assumed that the maximum number of factors is equal to the number of chatbot features listed in Table 4 (n = 21). As it is recommended that there be a minimum of three items per factor to produce a reliable solution (Costello & Osborne, 2005), three items were generated to capture each of the chatbot features listed in Table 4 in line with the definitions. Thus, the preliminary item pool comprised of 66 items.

Item generation followed recommendations listed in sources such as DeVellis (2016) such as avoiding double-barrelled items and exceptionally lengthy items that make it difficult for the respondent to comprehend the item. Items for a given chatbot feature were generated with useful redundancy, that is, they assess the same chatbot feature using different phrasings which is recommended during initial item testing such that the superior items can be selected and incorporated into the final scale.

The items in this questionnaire are to be rated on a 5-point Likert scale, which has been shown to result in higher quality data than those with more rating points (Revilla, Saris & Krosnick, 2014). At the beginning of the questionnaire, respondents are presented with the following prompt: "*Based on the chatbot you just interacted with, respond to the following statements*." Respondents are required to indicate the extent to which they agree with each statement using a rating scale from 1 to 5 (1 = *strongly disagree*, 2 = *somewhat disagree*, 3 = *neither agree or disagree*, 4 = *somewhat agree* and 5 = *strongly agree*). An odd number of response points are used as it is possible that participants may have genuinely neutral opinions about various chatbot features and should be allowed to express such opinions accurately when answering the questionnaire.

Each member of the research team generated items independently and the team convened to review the item pool together to ensure that the items were clear and reflected the relevant chatbot feature. Subsequently, the revised list of chatbot features (Table 4) and the preliminary item pool were reviewed in a series of focus groups, the activities for which are described in the next section.

3. Study 1: Focus Groups

3.1. Overview

The purpose of the focus groups was to obtain the opinions of end-users to assess the content adequacy associated with the list of chatbot features in Table 4 as well as gain feedback on the preliminary item pool. Specifically, we wanted to know: (a) if users understood each chatbot feature, (b) the extent to which users believed each feature to be relevant for satisfaction as well as why (or why not), (c) if the items were of good quality and if not, how to improve them and (d) if users could recognise which feature a given item was measuring. Overall, we wanted a better understanding of the features that contribute to user satisfaction with chatbots.

3.2. Methods

3.2.1. Participants.

16 students were recruited via SONA and convenience sampling at the University of Twente. The sample consisted of 8 males and 8 females (M_{age} = 22.1 years, SD_{age} = 2.84 years). Participants' nationalities were German (N = 6), Indian (N = 5), Bulgarian (N = 3) and Dutch (N = 2). Ten individuals listed psychology as their field of study while the remainder belonged to other fields such as industrial design and other engineering specialisations.

3.2.2. Procedure.

Before the participants arrive, the video camera is set up at the head of the table and adjusted once the participants arrive. Informed consent forms are placed on the table. The participants are seated around a rectangular table and the moderator is seated beside them. The assistant moderator is seated in the opposite corner of the room close to the camera to ensure its continuous operation and will also take handwritten notes in case of technical faults. The participants are greeted and briefly introduced to the study, their role for the session and a

rough timeline of how the session will progress. They are also informed that the session will be video recorded. If a participant does not want to be filmed, then we will ask them if they are okay with only their voices being recorded. If this is still not satisfactory, then the only material that will be recorded are the notes that the assistant moderator takes manually. After this short verbal introduction, they are given time to read and sign the informed consent forms in front of them.

Once informed consent has been obtained, they are asked to fill in a short demographic questionnaire and all the forms are collected by the assistant moderator. After this, the moderator guides the discussion as per the session script, deviating from the script when deemed potentially fruitful. The script is divided into three sections: (a) interactive demonstration, (b) feature review and (c) item review. In the interactive demonstration, the participants are given a basic definition of chatbots and a demonstration using the Finnair chatbot. The moderator operates the chatbot while asking participants to offer input for the chatbot so they can understand the fundamentals of a chatbot interaction. After the demonstration, the moderator asks participants to think about what they liked and did not like about the interaction as well as changes that they would like to see in it. At the feature review stage, participants are given the list of chatbot features obtained in the pre-experimental phase. They are given 15 minutes to mark beside each feature whether they thought it was relevant or not as well as a brief note to describe why they thought so if they could. Afterwards, the moderator resumes the discussion and asks the participants to bring up features that they believed to be very important and/or not at all important for them, opening the discussion to the other participants to voice their opinions. At the item review stage, participants are given the preliminary item pool also generated in the pre-experimental phase. They are given 15 minutes to mark beside each item which feature(s) the item is attempting to measure as well as

whether the item is clear or not. They are reminded that it is acceptable if they match an item to more than one feature as well as if an item cannot be matched to any feature at all.

3.2.3. Materials.

Informed consent. Participants are required to read and sign an informed consent form (Appendix 1.2) in which the study and the nature of the participant's contributions are described in as much detail as appropriate. Special attention was directed at briefing participants that they will be video recorded solely for data analysis purposes.

Demographics questionnaire. After obtaining informed consent, basic demographic information is acquired by asking the participant to fill in a brief form (Appendix 1.1). The demographic information collected comprises of: (1) gender, (2) age, (3) nationality, (4) field of study and (5) three questions related to prior experience with chatbots.

Session script. The research team collaborated with an expert to produce an appropriate script to guide the focus group session (Appendix 1.5). After deciding on the research goals that the focus group should meet, the team generated instructions and leading questions to ensure that the research goals were met. Extra questions and prompts were also generated in order to guide discussions in case the group requires a "push" in the right direction. Apart from the above-mentioned questions, text was inserted between the questions as deemed necessary in order to introduce or explain something for the participants. An expert was then approached for feedback on the script, whose recommendations were taken into consideration for the final script.

List of chatbot features. The list of chatbot features presented in Table were compiled into a document alongside their descriptions and printed out (Appendix 1.3). The list was split into two pages to make it easier for participants to read and process the entire list. Beside each

feature and its description is a column in which participants have to mark whether they believe the feature to be important or not to their satisfaction with information chatbots.

Preliminary item pool. The item pool generated for the list of 21 chatbot features and therefore comprising 63 items was compiled into a document and printed out (Appendix 1.4). The list was split into two pages to make it easier for participants to read and process the entire list. Beside each item was a column to mark which feature they believe the item measured and a second column to mark if they believed the item to be of good quality or not.

Video camera. A GoPro Hero 5 was used to video record the focus group discussions. This device was chosen for its ease of use and portability. Additionally, it can capture video in 4k resolution, providing high quality footage for further detailed analysis.

3.2.4. Data analysis.

Feature review. Participants were asked to indicate the relevance of each feature to their satisfaction with chatbot interactions on the list of features presented to them. Clear positive responses (e.g. ‘yes’, ‘very important’, ‘very relevant’, tick marks, etc.) were coded as “1” while all other response were coded as “0”. Clear positive responses were totalled for each feature and converted into a percentage score that indicated degree of consensus reached about the feature’s relevance to user satisfaction (Table 5).

All 21 features were classified into three categories (Table 6). Features that had a consensus score of 90% or more were classified as *very relevant*. Features that had a consensus score of less than 80% were classified as *unimportant*. Features for which consensus ranged between 80 to 90% were classified as *unclear*. All features were then reviewed further based on more qualitative data and expert review in order to determine whether a given feature should be retained or not.

Additionally, the research team reviewed the video footage of the focus groups. Specifically, we were interested in the specific factors that were raised during discussion, whether participants considered them relevant or not as well as their rationales for thinking so. The recordings were transcribed with enough detail to capture the above-mentioned details (Appendix 1.6). As participants were also told to write comments beside the relevant feature on the list presented to them, these comments were also compiled for optional qualitative reference.

Item review. Participants were asked to match each item in the item pool to the chatbot feature they believed the item measured. It was specified that an item can be matched to several features or none at all if the participant thought this was the case. If items were matched to the right chatbot feature, this was taken as evidence of content validity and item quality. Items matched to more than one feature were marked as potentially problematic (Appendix 1.7). Qualitative comments on specific items that were expressed during the focus group discussions were also compiled (Appendix 1.6).

3.3. Results

The results will cover two main points. The first section will present the results regarding the refinement of the list of chatbot features presented in Table 4. The second section will present a qualitative interpretation of the USIC construct in light of the list of chatbot features that were retained as important determinants of user satisfaction with information chatbots.

Table 5*: Consensus ratings for list of 21 chatbot features (in descending order)*

	Chatbot feature	Consensus (%)		Chatbot feature	Consensus (%)
1.	Response time	*100*	***12.***	Ability to maintain themed discussion	*88*
2.	Perceived credibility	*100*	***13.***	Maxim of relation	*87*
3.	Understandability	*100*	***14.***	Trust (general)	*81*
4.	Maxim of quantity	*100*	***15.***	Appropriate language style	*80*
5.	Ease of use (general)	*100*	***16.***	Process tracking	*80*
6.	Expectation setting	*100*	***17.***	Ease of starting a conversation	*79*
7.	Flexibility of linguistic input	*94*	***18.***	Engage in on-the-fly problem solving	*73*
8.	Reference to service	*94*	***19.***	Graceful responses in unexpected situations	*69*
9.	Privacy & security	*93*	***20.***	Personality	*50*
10.	Visibility (website only)	*93*	***21.***	Enjoyment	*50*
11.	Recognition and facilitation of user's goal and intent	*93*			

Table 6*: List of 21 chatbot features classified into three categories based on consensus ratings*

Very relevant	Unclear	Unimportant
Response time	Ability to maintain themed discussion	Personality
Perceived credibility	Maxim of relation	Enjoyment
Understandability	Trust (general)	Graceful responses in unexpected situations
Maxim of quantity	Appropriate language style	Engage in on-the-fly problem solving
Ease of use (general)	Process tracking	Ease of starting a conversation
Expectation setting		
Flexibility of linguistic input		
Reference to service		
Privacy & security		
Visibility		
Recognition and facilitation of user's goal and intent		

3.3.1. Refinement of list of chatbot features.

Excluded features.

Ease of Use. As expected, *ease of use* was essential for many participants - as one participant expressed, “if it’s not easy to use, then I would never use it again”. It was appreciated if chatbots gave clear indications and instructions on how to interact with it, “essentially guiding the user through the process”. Participants did, however, agree that the general *ease of use* feature was too vague and could mean different things and as such, they would find it difficult to respond the corresponding items. Encouragingly, many participants also noticed the relationships between *ease of use* and the specific features that captured different aspects of ease of use (e.g. flexibility of linguistic input, ease of starting a conversation, understandability, expectation setting, etc.) Consistent with the rationale, participants believed it is important that the specific features pertaining to *ease of use* be retained as the specific features did indeed capture different but relevant aspects of the broader construct. This therefore rendered the general *ease of use* feature redundant and was therefore removed.

Trust. Like the *ease of use* feature, *trust* was also considered to be important when interacting with a chatbot. Several participants felt strongly about their reluctance to reveal personal information to the chatbot in the process of communicating their request and therefore it was unsurprising that *trust* was largely interpreted alongside one of its specific features *privacy and security.* Another specific feature that was devised to represent another aspect of trust was *perceived credibility* and while this was not as frequently associated with trust, participants were consistently in agreement that a trustworthy chatbot should provide answers that are or at least appear to be “true and based on fact”. Several participants found the general *trust* feature to be “subject to interpretation” and asked for further clarification as to what trust

refers to in this context, providing support for the retention of the specific features *privacy & security* and *perceived credibility* and the exclusion of the broad counterpart.

Personality and Enjoyment. *Personality* and *enjoyment,* while "nice to have", were largely regarded as irrelevant and even unnecessary in information-retrieval chatbots. This is consistent with the elimination of personality as an unimportant feature by Tariverdiyeva and Borsci (2019) and provides strong evidence that the chatbot's personality truly does not play an important role in shaping USIC. While most participants did notice when a chatbot provided a fun, engaging interaction such as the demonstration with Finnair's chatbot Finn who was perceived to be "sweet" and had a "good attitude", their comments revealed that their primary motivation for engaging with such chatbots is to obtain relevant information quickly and with ease. Consistent with this motivation, several participants said that they do not expect an information-retrieval chatbot to be 'humanlike' because "it's a robot, it's not a human". Of interest is the observation that despite being distinct features, participants generally talked about these two features together, suggesting a close relationship between chatbot personality and the level of fun or enjoyment experienced. Consistent with some studies, it was suggested by some that being fun and likeable may be more important for chatbots designed for other specific purposes such as health or entertainment (Fadhil, 2018; Fadhil & Schiavo, 2019).

Process Tracking. Many participants did not understand why the feature *process tracking* should be at all relevant for information-retrieval chatbots. It was common understanding that responses to any request are delivered almost immediately - in fact, the quick nature of chatbot responses if one of the main reasons users appreciate and would use such chatbots. If this is the case, participants were confused as to what "process" is occurring that is taking enough time that the user needs to be kept informed about instead of being given the response immediately – "it just needs to give you direct answers when [you] ask". This feature was therefore excluded.

Appropriate Language Style. Participants believed the language style did not matter as long as they understood the chatbot's responses (see feature: *understandability*). The research team also concluded that as long as the language style is not impolite or derogatory, this feature becomes irrelevant for determining satisfaction. The team also came to the conclusion that while not relevant for the measure currently being developed, the use of a minimally appropriate language style is a feature that should be considered by designers consistent with comments from certain participants who agreed that like personality, language style must be mildly appropriate to the service the chatbot provides and gave the example of a chatbot that provided funeral home services.

On-the-Fly Problem Solving. Almost all participants had trouble distinguishing between the features *on-the-fly problem solving* and *response time*. A participant explained that [she thinks] "a chatbot is supposed to help [you] solve a problem and not actually solve the problem itself". One participant offered an explanation that allowed us to clarify these two features. Specifically, he saw *on-the-fly problem solving* as composed of two components: the capability of the chatbot to adequately address the user's request for information and doing so immediately with minimal delay. While *response time* refers to the latter component and was therefore retained, the former component is captured by other features in the list thus eliminating the need for this feature. Additionally, *response time* was renamed to *perceived speed* to be more consistent with user perception of the chatbot's response speed rather than an objective measure of the same.

Retained features.

This section describes rationale underlying the retention of certain features despite quantitative data suggesting that these factors were irrelevant.

Graceful responses to unexpected situations. This factor scored very low on relevance, but the factor was retained nevertheless. Participants had voiced in discussion that they would like chatbots to be able to "sense when they can't help [me]" and provide a way forward for the user to accomplish their goal instead of breaking down communication altogether, such as repeatedly producing irrelevant output or showing the same error message. Participants did not seem to understand the description when individually indicating their responses on paper but upon explanation and discussion afterwards, agreed that it was important to their interaction. In response to the participants' comments, the feature was renamed to *graceful breakdown.*

Ease of starting a conversation. This factor also scored low on relevance based on the ratings provided by the participants. When asked about why they rated this factor to be unimportant, many participants voiced their confusion about how this was different from another factor *visibility.* Specifically, they believed that if the chatbot was visible and easy to access, how hard could it be to start a conversation with it? However, once the intended meaning of the factor was explained, several participants expressed their understanding that while the two factors might be related but are still distinct and important. In fact, one participant recounted his experience of how he could not get a certain chatbot to understand what he wanted for such a long time that he gave up and would have appreciated some advice or options at the beginning to make it easier. While many participants agreed that this feature was important, some did mention that users who have had significant experience with chatbots and/or are willing to be patient will have less of a struggle in this aspect. Given the above points, this feature was retained.

Furthermore, there were several features that participants rated as irrelevant because they did not understand after reading the initial feature descriptions, the meanings for which only became clearer once the moderator clarified and explained each feature in greater detail. Considering such misunderstandings, renaming and rewording of feature descriptions took place for other chatbot features in order to better capture the intended meaning of the feature, namely *visibility*, *flexibility of linguistic input*, *privacy & security* and *maxim of relation.*

3.3.2. Feedback on item pool.

Overall, the item pool received positive feedback along item quality. Items were deemed to be well-worded, clear and easy to understand. All items were matched to the right chatbot feature. However, several items were found to be perceived as assessing multiple chatbot features in addition to the right one which indicates that item quality can be improved by rewording these items appropriately (Appendix 1.7). On the topic of useful redundancy, one participant commented that many items assessing the same chatbot feature appear quite similar and would benefit from rephrasing. Given the general positive feedback on the item pool and the time constraints faced while conducting this study, it was decided to proceed without the further refinement of items.

Table 7 shows the revised list of 14 chatbot features that were obtained based on the results of the focus groups along with reworded descriptions in italics and reworded feature names with asterisks (*).

3.4. Limitations

The findings from these focus groups may have limited generalizability to other end-user groups as the sample was relatively homogenous. Specifically, the participants were all university students with above average proficiency with the English language and a mean age of 22 years. This sample represents one of the major end-user groups of chatbots, that is the

population that is relatively young, tech-savvy and experienced with the use of instant messaging and other emerging technologies. It may be the case that a different set of chatbot features may prove more significant in determining user satisfaction for other end-user groups such as the elderly and novices. Further research must be conducted to explore if the same set of features underlies satisfaction for all potential end-user groups.

There is the possibility that some important points were not raised during the focus group discussions because of inadequately lively discussions during some sessions. Given the time constraints and moderately inexperienced moderators, it is possible that a conducive enough atmosphere for discussion was not created. However, as fruitful discussions were also observed in other sessions, perhaps more attention should be paid to the way in the participants are grouped together. Additional effort should be invested in writing detailed session scripts to address potentially inadequate participation in focus groups by referring to the relevant literature, such as the guidelines provided by Krueger and Casey (2002) on how best to encourage group participation in focus groups.

Finally, it was noticed that there was a discrepancy between the individual ratings that participants made on their papers and the opinions raised by those individuals during the discussion. On reflection, the discrepancy can be attributed to the fact that for many participants, descriptions of several chatbot features was unclear when they were rating each feature individually. As the descriptions were clarified during discussion, their opinions changed and their ratings, which were done beforehand, were not always consistent with their changed perspectives. Indeed, upon realising this, we relied more on the opinions voiced during the discussion about different chatbot features and used feedback on these misunderstandings to refine the chatbot feature names and descriptions.

Table 7*: Revised list of 14 chatbot features after focus groups (in no particular order)*

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

3.5. Qualitative interpretation of the USIC construct.

While all the chatbot features could not be discussed in great depth given time constraints, participants were asked to bring up in discussion the features they believed to be the most important in an interaction with a chatbot. Over the course of the focus group discussions, a set of chatbot interaction characteristics emerged across all groups as essential. End-users ultimately want the chatbot to be able to *facilitate the accomplishment of their goal* and this means several things. The qualitative interpretation below follows the communication flow between the end-user and the chatbot.

Unfortunately, the interaction ends before it even begins if initiating a conversation with the chatbot becomes a daunting task – sometimes, it simply cannot be found, either on a messaging platform or a website, but a more common issue is when users are faced with a somewhat blank dialogue boxes, uninformative prompts and confusion regarding how to proceed further. Thoughts such as “do I just start typing?”, “is there a button to click somewhere?” and “what exactly do I say?” are common and make it less likely that the user continues the interaction. More experienced users seem to understand that the technology underlying most chatbots leverages on simple keyword associations that merely need to be typed into the box, but this is not the case for neither all users nor all chatbots. Instead, participants mentioned that they appreciate minimal but informative indications that guide the user towards a smoother interaction. It appears that *accessibility*, *ease of starting the conversation* and *expectation setting* play essential roles in making it easy for the user to begin interacting with the chatbot. However, frustration ensues if the chatbot cannot understand the user’s request for information. Users expect that the chatbot be ‘intelligent’ enough to understand the user regardless of how the request has been put across and perceive the chatbot as highly incompetent if it cannot do so. This includes, for example, choice of words, spelling and grammatical mistakes, missing words and even extremely long or short sentences. Users

"don't want to repeat [themselves]" and "think too much" about how best to phrase their request. The *communication effort* involved that users expend into making the chatbot understand their request appears to be crucial to user satisfaction.

Once the user has been able to communicate their request successfully to the chatbot, then the focus shifts to the chatbot's responses. Participants are visibly irritated when they do not receive a "helpful" or a "useful" response which, upon clarification, generally refers to a response that is at least moderately relevant to the user's goal. However, *relevance* was also interpreted in conjunction with other chatbot characteristics. For example, items associated with the *perceived credibility* or accuracy of the information were also matched to features such as relevance and ability to maintain *themed discussion*. Indeed, a response that is relevant to the request and context is also likely to be perceived as accurate and precise in that it is providing the information that the user finds helpful. Additionally, helpful, relevant responses also crucially include *reference to the service* such as relevant hyperlinks or automatic transitions as these present the user with the choice of obtaining more information. This relates closely to another characteristic that was co-mentioned with relevance: *maxim of quantity*. Participants mentioned that "only the relevant information" should be presented in the dialogue box – the chatbot's response should "get to the point" and make sure it does not give "too much and unnecessary information". A frequent complaint that emerged during the chatbot demonstration was that the chatbot initially presented too much information, resulting in information overload even though it was attempting to help. If there is too little or too much information, it becomes markedly difficult for the user to make sense of the response even if it contains the information they want. However, there are characteristics apart from quantity that determine the *understandability* of the chatbot's response such as the language style, the complexity of the words used, the way the information has been structured, etc. Of vital importance is the ability for the chatbot to "recognise when it can't help [me]" and exhibit

graceful breakdown. Instead of breaking down in ways such as repeatedly showing the same error message or consistently responding with irrelevant information, participants expect it to let the user know that "it can't help [you] nicely" and act accordingly, such as providing a link to customer service contact information. Additionally, almost every participant expected that the chatbot respond rapidly to their request (*perceived speed*) as this was mentioned to be one of the main advantages of chatbots over other ways of obtaining information. Finally, it was important that the whole interaction feel secure. *Perceived privacy* is an important issue in a time when it is difficult to ensure that one's personal information is kept secure and this becomes even more relevant in the context of chatbots that are embedded on social media platforms such as Facebook which have access to a wealth of private data.

In the aftermath of Study 1, 14 out of 21 chatbot features remained as essential to user satisfaction with information chatbots. What was striking was the small number of features considered truly irrelevant - it appears that end-users appear to hold information-retrieval chatbots to significantly high standards. The reason for this becomes clear upon discovering that in addition to productivity (Brandtzaeg & Folstad, 2017; 2018; Folstad & Skjuve, 2019), a significant driver of adoption for information chatbots is the chatbot's superiority over the alternatives (e.g. smartphone applications, websites, search engines) that end-users are more familiar with for the purpose of retrieving information (Beriault-Poirier, Tep & Senecal, 2018; Zamora, 2017; Abu Shawar & Atwell, 2016). In short, the findings suggest that satisfaction with information chatbots is largely determined by the chatbot's absolute effectiveness and efficiency in helping the user accomplish their goal but also relative to existing alternatives. As such, all the chatbot features consistent with the above expectations were considered important, accounting for the composition of the revised list in Table 7. While other studies that have largely reported low to moderate correlations between effectiveness, efficiency and satisfaction for other interfaces, the notion that these three constructs may be more closely related in the

case of information-retrieval chatbots should be investigated in future research. Ultimately, end-user expectations indicate the presence of a relatively high adoption barrier that information-retrieval chatbots need to overcome in order to ensure that users are satisfied and engage in continued usage.

3.6. Towards a theoretical model of USIC

The refined list of chatbot features in Table 7 were clustered on the basis of the qualitative interpretation of USIC presented in the previous section as well as specific comments raised by participants during the focus groups. This resulted in two hypothesised models of USIC which are presented below, comprising eight (Table 8) and five (Table 9) factors respectively.

Table 8*: Proposed 8-factor structure of USIC*

	Factor	Description	Items
1	Initiating conversation	How easy it is for the user to start interacting with the chatbot, including not only accessibility but also how simple it feels to actually start the conversation i.e. to start typing	Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8, Y9
2	Communication effort	How easy it is for the user to successfully (or not) convey his or her information-retrieval goal to the chatbot	Y10, Y11, Y12
3	Content relevance	The extent to which the chatbot’s response addresses the user's request	Y13, Y14, Y15, Y22, Y23, Y24, Y25, Y26, Y27, Y37, Y38, Y39
4	Response clarity	How easy it is for the chatbot’s response to be understood by the user	Y34, Y35, Y36, Y28, Y29, Y30
5	Reference to service	The ability of the chatbot to provide useful and relevant hyperlinks or automatic transitions either in lieu of or in addition to its response to the user's request	Y16, Y17, Y18

6	Graceful breakdown	The appropriateness of the manner in which the chatbot responds if and when it encounters a situation in which it cannot help the user	Y31, Y32, Y33
7	Perceived speed	How quickly the chatbot responds to each input the user gives	Y40, Y41, Y42
8	Perceived privacy	How secure the entire interaction feels as a consequence of revealing potentially personal information to the chatbot	Y19, Y20, Y21

Table 9*: Proposed 5-factor structure of USIC*

	Factor	Description	Items
1	Communication quality	How easy it is for the user to communicate his or her information-retrieval goal	Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8, Y9, Y10, Y11, Y12
2	Response quality	The overall quality of the chatbot's response once the user has provided some form of input to the chatbot	Y13, Y14, Y15, Y16, Y17, Y18, Y22, Y23, Y24, Y25, Y26, Y27, Y28, Y29, Y30, Y31, Y32, Y33, Y37, Y38, Y39, Y34, Y35, Y36,
3	Graceful breakdown	The appropriateness of the manner in which the chatbot responds if and when it encounters a situation in which it cannot help the user	Y31, Y32, Y33
4	Perceived speed	How quickly the chatbot responds to each input the user gives	Y40, Y41, Y42
5	Perceived privacy	How secure the entire interaction feels as a consequence of revealing potentially personal information to the chatbot	Y19, Y20, Y21

4. Study 2: Questionnaire Evaluation

4.1. Overview

At the end of Study 1, the list of chatbot features that were considered important determinants of user satisfaction was refined and the preliminary item pool was reviewed by potential end-users. As the list of features was reduced from 21 to 14 features, the new preliminary item pool now consists of 42 items. The item pool will be administered to a sample of potential end-users as a post-test questionnaire to measure level of user satisfaction with various chatbots. The resulting dataset and two previously devised theoretical models (Tables 8 and 9) will be used to perform factor analysis to confirm or uncover the underlying factor structure and investigate the evidence in support of the validity and reliability of the preliminary questionnaire.

4.2. Methods

4.2.1. Participants.

60 students were recruited via convenience sampling at the University of Twente. The sample consisted of 37 males and 22 females (M_{age} = 23.7, SD_{age} = 4.80). Participants' nationalities comprised of Dutch (5%), German (44%) and others (51%). 19 participants listed psychology as their field of study while the remainder belonged to other fields such as engineering and business administration.

4.2.2. Procedure.

Participants are invited into the usability testing room. After being given a brief introduction about the study, they are instructed to read and sign the informed consent form placed on the table. After obtaining informed consent, the researcher opens the Qualtrics survey that has been designed for this experiment on the computer and gives the participant an

overview of the session. Specifically, participants are told that they will be assigned five random chatbots and are given an information-retrieval task to complete using each chatbot after which they will respond to a 42-item questionnaire about their experience with the chatbot. Once the participant is ready to begin the session, the researcher initiates screen recording. The participant is told to follow the instructions presented to them on the Qualtrics survey and ask the researcher if anything is unclear.

4.2.3. Materials.

Informed consent. Participants are required to read and sign an informed consent form (Appendix 1.2) in which the study and the nature of the participant's contributions are described in as much detail as appropriate.

Demographics questionnaire. After obtaining informed consent, basic demographic information is acquired by asking the participant to fill in a brief form (Appendix 1.1). The demographic information collected comprises of: (1) gender, (2) age, (3) nationality, (4) field of study and (5) three questions related to prior experience with chatbots.

Chatbots and tasks. Ten chatbots will be tested in this study, each of which will be associated with an information retrieval task for users to complete using the chatbot (Appendix 2.1). Four of these ten chatbots have been taken from a previous study (Tariverdiyeva & Borsci, 2019), therefore the task for these chatbots was taken from that study as well. The research team also collected six new, viable information-retrieval chatbots which made use of either websites or Facebook Messenger as the delivery platform. For each of the new chatbots, new tasks were devised. Tasks were generated by exploring each relevant website to discover possible information that users may request from the service through the chatbot and scenarios were generated surrounding these tasks to provide users with a use case.

Preliminary item pool. 42 items that capture the retained features were compiled into a questionnaire (Appendix 1.4). Participants respond to each item based on the extent to which they agree with the statement after interacting with the relevant chatbot using a 5-point Likert scale from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”).

Qualtrics survey. The experiment will largely be delivered by means of a Qualtrics survey which comprises of several sections (Appendices 2.3 and 2.4). For a given chatbot, the first section introduces each chatbot by the name of the business it serves (e.g. Booking.com) and a link that takes the participant to the relevant website where the chatbot can be found. Participants are told to copy the link into a new tab on the web browser and access the chatbot. The hyperlink has been removed in order to ensure that participants do not directly click on the link and lose the page on which they are completing the questionnaire. The next section presents the task that the participant performs with the chatbot. Participants are told to stay on this slide until they have decided that they have completed the task in case they need to refer to the task later. The third section presents the item pool that participants are instructed to respond to in order to assess the chatbots they just interacted with. The researcher will moderate the session guided by the session script (Appendix 2.2).

Each participant only tests five of ten chatbots. The assignment of these chatbots depends on a special randomisation procedure described below. The ten chatbots consist of four pre-tested chatbots and six new chatbots. Each participant will test two chatbots of known usability and three chatbots of unknown usability. Participants are randomly assigned to one of two conditions – in condition A, two pre-tested chatbots (one good chatbot and one bad chatbot) are provided and in condition B, the other two pre-tested chatbots (one good chatbot and one bad chatbot) are provided. Each participant is also randomly assigned to test three of the remaining six new chatbots. Additionally, the order of all five presented chatbots is randomised. The randomisation procedure is carried out through the Qualtrics survey software

randomisation tool. The benefits of doing so include reducing the time needed to administer and code the protocol, and more importantly, reducing the burden placed on each participant. By doing so, the resulting data may be more valid, stronger effects may be observed and participants may respond to more items (Little & Rhemtulla, 2013).

4.2.4. Data analysis.

Data preparation. Data cleaning and preparation was performed using Microsoft Excel. Data for three participants was discarded due to missing data. This missing data resulted from technical difficulties encountered when testing certain chatbots as server errors were present on occasion during the period of testing. The analysis was intended to be performed on a design × item dataset, which would have been obtained by averaging scores for each item across all participants for each chatbot design. However, due to pragmatic reasons i.e. limited number of chatbot designs tested, the analysis was instead conducted using a person × item dataset, which was obtained by averaging scores for each item across chatbot designs for each participant. Further explanation can be found in the results and discussion sections. Thus, for the data of the remaining 57 participants, responses for each item on the questionnaire were averaged across all tested chatbots for every participant to produce a mean score for each item on the questionnaire. The resulting data file was imported into R studio for analysis.

Confirmatory factor analysis. As two models, comprising eight and five factors respectively, were devised (Tables 8 and 9), it was decided to test these two models by performing confirmatory factor analysis. We use a Bayesian confirmatory factor analysis (bCFA) algorithm developed by Merkle and Rosseel (2015) as part of the R package ‘blavaan’. The parameters of M1 and M2 were specified according to model 1 (Table 8) and model 2 (Table 9) respectively. The code used to run this analysis can be found in Appendix 2.6. Neither model converged after a certain number of iterations despite adjustments made to the

arguments. The lack of convergence could be due to several reasons, including but not limited to an insufficient sample size, insufficient MCMC chain length or inaccurately hypothesised models. As the bCFA was unsuccessful, it was decided to perform an exploratory factor analysis to obtain a factor structure that best represented the data.

Parallel analysis. Prior to the exploratory factor analysis, a parallel analysis was conducted in order to determine the number of factors to extract as this is argued to be one of the more accurate factor retention methods (Hayton, Allen & Scarpello, 2004). Parallel analysis was conducted using the *fa.parallel* function from the R package 'psych' (Revelle, 2017). The number of factors to retain is indicated by where the tracings for actual (blue line) and simulated data (red line) cross such that factors that lie above the crossing exhibit eigenvalues with magnitudes that are greater than would be expected by chance alone and thus should be retained. The code used for this analysis can be found in Appendix 2.6.

Exploratory factor analysis. Bayesian exploratory factor analysis was performed using an algorithm developed by Conti et al. (2014) that can be found in the R package 'BayesFM'. The sampler was run with a burn-in period of 5000 followed by 50000 iterations for posterior inference. The maximum number of factors to extract was informed by the results of the parallel analysis. After MCMC sampling is complete, MCMC draws were processed a posteriori to solve the sign and column switching problem that impedes the interpretation of the indicator matrix. The code used for this analysis can be found in Appendix 2.6.

Reliability analysis.

Reliability analysis was conducted individually for each latent factor using the alpha function from the R package 'psych' (Revelle, 2017). The code used for this analysis can be found in Appendix 2.6.

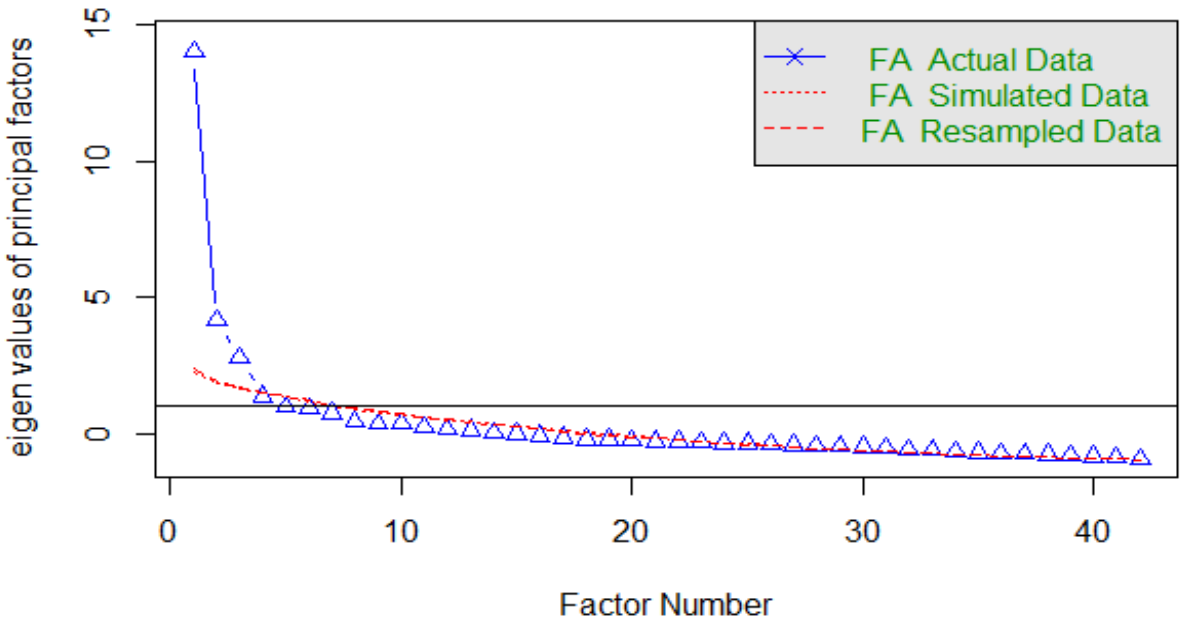
4.3. Results

The following section presents the results of (a) parallel analysis, (b) exploratory factor analysis and (c) item evaluation and selection procedures. As stated previously, all analyses were performed on a person × item dataset. The obtained results are thus psychometric in nature and describe if and how *individuals* differ in their responses to the item pool and therefore reflect individual differences in how different aspects of information chatbot interactions are evaluated. However, the desired outcome would be to describe if and how *chatbot designs* differ along user satisfaction based on the scores obtained for each item and the relevant analysis must be performed on a design × item dataset. These results therefore represent a first approximation of the factor structure underlying this questionnaire based on the data collected, in which the object of measurement is the individual rather than the chatbot itself.

4.3.1. Parallel analysis.

The results of the parallel analysis are shown in Figure 3, which suggests that the number of factors to extract is three. However, it has been cautioned to use parallel analysis in conjunction with other factor retention methods under certain conditions, including but not limited to small sample sizes and highly correlated factors (Turner, 1998). Subsequent visual inspection of the ‘elbow’ in the scree plot generated by the actual data, represented by the blue line in Figure 3, suggests that the number of factors may lie between 3 and 7.

Figure 3*: Parallel analysis scree plots*



4.3.2. Factor retention.

Factor solution 1 (7 factors).

Taking a more conservative approach, the maximum suggested number of factors was extracted first (k=7). The posterior probability of the highest probability model is 0.173 (k=5). Posterior means of factor loadings are presented in Table 13. Almost all factor loadings show significantly high magnitudes, indicating that most items strongly load onto their respective dedicated factors. Factor loadings for items 19, 20 and 21, which load onto Factor 4 and clearly assess perceived privacy, exhibit significantly lower magnitudes of posterior means and are accompanied by significant uncertainty in their true magnitude as observed from the credibility intervals, indicating high uncertainty associated with this factor. Factor 3, comprising of items 7, 8, 9, 13 and 14, was found to be difficult to interpret in a meaningful manner. It is therefore unlikely that this structure best represents the data and captures the construct at hand.

Factor solution 2 (4 factors).

Subsequently, a structure comprising four factors was extracted. The posterior probability of the highest probability model is 0.699 (k=4). Posterior means of factor loadings are presented in Table 14. All items exhibit significantly high factor loadings and are accompanied by strong 95% credibility intervals, indicating that all items strongly load onto their dedicated factors. Factor interpretations are presented in Table 12. The latent factors can be interpreted meaningfully and coherently. Correlations between latent factors can be found in table 10. Correlations indicate high multicollinearity between latent factors which points to a lack of discriminant validity between the factors.

Table 10*: latent factor correlations for factor solution 2 (k=4)*

	F1	**F2**	**F3**	**F4**
F1	1			
F2	0.989	1		
F3	0.971	0.975	1	
F4	0.992	0.987	0.966	1

Factor solution 3 (3 factors).

Finally, a structure comprising three factors was extracted. The posterior probability of the highest probability model is 0.604 (*k*=3). Posterior means of factor loadings are presented in Table 15. All items exhibit significantly high factor loadings and are accompanied by strong 95% credibility intervals, indicating that all items strongly load onto their dedicated factors. Correlations between latent factors can be found in Table 11. Correlations indicate high multicollinearity between latent factors which points to a lack of discriminant validity between the factors. Factor interpretations are presented in Table 12. To a large extent, the latent factors can be interpreted meaningfully and coherently. While items 40, 41 and 42, which assess

perceived speed, loaded onto a separate factor in factor solution 2 (k=4), these items were found to load onto Factor 1 alongside other items that assess communication quality. As perceived speed is qualitatively distinct from communication quality, it should be retained as a separate factor.

Table 11*: latent factor correlations for factor solution 3 (k=3)*

	F1	**F2**	**F3**
F1	1		
F2	0.990	1	
F3	0.973	0.977	1

A comparison of factor interpretations for factor solutions 2 and 3 can be found in table 12, in which italicised text is used to highlight the main differences. Empirically, factor solutions 2 (*k*=4) and 3 (*k*=3) are almost identical, although the highest probability model for factor solution 2 is marginally higher. Factor solution 3 includes the chatbot feature pertaining to perceived speed under factor 1 which otherwise comprises of chatbot features relevant to the relative ease or difficulty with which the user communicates their request to the chatbot, making it difficult to interpret the factor. On the other hand, factor solution 2 places perceived speed as a separate factor, which is more meaningful for interpretation. It appears that factor solution 2, comprising four factors, not only fits the data but is also consistent with the qualitative findings from the focus group discussions and will be selected for subsequent analysis.

Table 12*: Comparison between factor interpretations for factor solutions 2 and 3*

	Factor solution 2 (k=4)				***Factor solution 3 (k=3)***		
	Items	**Chatbot Features**	**Factor Interpretation**		**Items**	**Chatbot Features**	**Factor Interpretation**
F1	Y1, Y2, Y3, Y4, Y5, Y6, Y10, Y11	Accessibility, Ease of Starting a Conversation, Flexibility of Linguistic Input	Communication Quality	**F1**	Y1, Y2, Y3, Y4, Y5, Y6, Y10, Y11, Y40, Y41, Y42	Accessibility, Ease of Starting a Conversation, Flexibility of Linguistic Input, *Perceived Speed**	Communication Quality
F2	Y7, Y8, Y9, Y12, Y14, Y15, Y16, Y17, Y18, Y22, Y23, Y24, Y25, Y26, Y27, Y28, Y29, Y30, Y31, Y32, Y33, Y34, Y35, Y36, Y37, Y38, Y39	Expectation Setting, Themed Discussion, Reference to Service, Recognition and Facilitation of Goal & Intent, Maxim of Relation, Maxim of Quantity, Graceful Breakdown, Understandability, Perceived Credibility, Communication Effort	Response Quality	**F2**	Y7, Y8, Y9, Y12, Y13, Y14, Y15, Y16, Y17, Y18, Y19, Y20, Y21, Y22, Y23, Y24, Y25, Y26, Y27, Y28, Y29, Y30, Y31, Y32, Y33, Y34, Y35, Y36, Y37, Y38, Y39	Expectation Setting, Themed Discussion, Reference to Service, Recognition and Facilitation of Goal & Intent, Maxim of Relation, Maxim of Quantity, Graceful Breakdown, Understandability, Perceived Credibility, Communication Effort	Response Quality
F3	Y13, Y19, Y20, Y21	Themed Discussion, Perceived Privacy	Perceived Privacy	**F3**	Y13, Y19, Y20, Y21	Perceived Privacy	Perceived Privacy
F4	*Y40, Y41, Y42*	*Perceived speed*	*Perceived Speed**				

Table 13: *Posterior means of factor loadings for factor solution 1 (k=7)*

	Posterior means of factor loadings					95% credibility interval	
Item (Y)	**F1**	**F2**	**F3**	**F4**	**F5**	**Lower**	**Upper**
1	4.170					3.435	4.973
10	3.036					2.486	3.643
11	3.229					2.636	3.869
2	4.109					3.397	4.903
3	4.184					3.456	4.970
4	4.137					3.409	4.930
5	3.860					3.203	4.607
6	3.967					3.297	4.745
12		3.526				2.933	4.258
15		3.319				2.714	3.933
16		3.670				3.063	4.416
17		3.840				3.178	4.609
18		3.862				3.204	4.630
22		3.303				2.720	3.946
23		3.606				3.007	4.337
24		3.404				2.793	4.038
25		3.286				2.719	3.930
26		3.349				2.795	4.023
27		3.394				2.800	4.043
28		3.283				2.689	3.906
29		3.370				2.785	4.029
30		3.059				2.538	3.677
31		2.841				2.328	3.391
32		3.087				2.533	3.708
33		3.130				2.569	3.726
34		3.585				2.977	4.289
35		3.697				3.029	4.403
36		3.847				3.174	4.595
37		3.476				2.867	4.144
38		3.620				2.987	4.315
39		3.726				3.079	4.436
7		3.695				3.069	4.439
13			2.872			2.352	3.460
14			3.185			2.600	3.773
8			3.374			2.781	4.026
9			3.542			2.901	4.205
19				0.375		-3.679	3.748
20				0.300		-3.024	3.015
21				0.355		-3.457	3.581
40					-4.282	-5.130	-3.573
41					-4.294	-5.126	-3.553
42					-4.258	-5.050	-3.484

Table 14*: Posterior means of factor loadings for factor solution 2 (k=4)*

	Posterior means of factor loadings				95% credibility interval	
Item (Y)	**F1**	**F2**	**F3**	**F4**	**Lower**	**Upper**
1	4.174				3.434	4.972
10	3.037				2.463	3.624
11	3.234				2.630	3.864
2	4.114				2.896	4.204
3	4.191				2.335	3.448
4	4.141				2.592	3.783
5	3.864				2.711	3.943
6	3.971				3.026	4.372
12		3.527			3.146	4.573
14		3.179			3.173	4.604
15		3.323			2.598	3.775
16		3.674			3.370	4.874
17		3.844			2.088	3.069
18		3.864			2.459	3.575
22		3.306			2.717	3.934
23		3.606			2.954	4.286
24		3.408			2.781	4.030
25		3.290			2.726	3.934
26		3.353			2.773	4.005
27		3.396			2.784	4.023
28		3.286			2.698	3.929
29		3.371			2.773	4.016
30		3.060			3.443	4.984
31		2.844			2.502	3.633
32		3.090			2.322	3.385
33		3.133			2.514	3.685
34		3.589			2.549	3.718
35		3.701			2.945	4.266
36		3.851			3.031	4.409
37		3.476			3.157	4.574
38		3.624			2.842	4.122
39		3.729			2.995	4.327
7		3.699			3.072	4.446
8		3.356			3.414	4.925
9		3.517			3.538	5.107
13			2.875		3.556	5.127
19			3.176		3.526	5.086
20			2.567		3.179	4.591
21			3.003		3.254	4.705
40				4.288	3.046	4.406
41				4.299	2.724	3.985
42				4.264	2.869	4.194

Table 15*: Posterior means of factor loadings for factor solution 3 (k=3)*

	Posterior means of factor loadings			95% credibility interval	
Item (Y)	**F1**	**F2**	**F3**	**Lower**	**Upper**
1	4.173			3.441	4.978
10	3.037			2.473	3.628
11	3.233			2.616	3.858
2	4.115			3.380	4.892
3	4.193			3.436	4.976
4	4.143			3.417	4.943
40	4.277			3.508	5.086
41	4.282			3.498	5.077
42	4.248			3.467	5.030
5	3.862			3.178	4.602
6	3.970			3.262	4.724
12		3.527		2.905	4.220
14		3.180		2.599	3.796
15		3.323		2.715	3.948
16		3.675		3.007	4.372
17		3.844		3.141	4.575
18		3.865		3.174	4.609
22		3.306		2.703	3.928
23		3.607		2.969	4.299
24		3.408		2.788	4.046
25		3.290		2.688	3.899
26		3.354		2.773	4.005
27		3.397		2.790	4.037
28		3.286		2.705	3.938
29		3.372		2.758	4.009
30		3.061		2.513	3.653
31		2.845		2.336	3.406
32		3.091		2.529	3.693
33		3.133		2.560	3.733
34		3.590		2.944	4.272
35		3.701		3.026	4.398
36		3.852		3.165	4.590
37		3.476		2.839	4.131
38		3.624		2.967	4.313
39		3.729		3.047	4.427
7		3.699		3.054	4.425
8		3.355		2.745	4.008
9		3.517		2.863	4.202
13			2.874	2.325	3.435
19			3.177	2.604	3.780
20			2.566	2.088	3.072
21			3.004	2.466	3.574

4.3.3. Item evaluation and selection.

In order to evaluate the items, descriptive statistics and corrected item-total correlations were computed for each item (Appendix 2.5). Histograms for responses associated with each item can be found in Appendix 2.6. Items with means greater than 4 were flagged as item means that are distant from the midpoint of the Likert scale can indicate a lack of variance in the responses. The flagged items were found to belong to the factors communication quality and perceived speed. Additionally, two items from communication quality were noted to have corrected item-total correlations with values less than 0.3, which suggests that they currently do not correlate well with the sub-scale. There were no other significant findings at this stage. With this empirical data in mind, item selection was subsequently conducted.

There were two qualitative principles upon which item selection was performed. First, we wanted a short yet useful measurement tool therefore the intent was to drop ‘bad’ items in order to optimise scale length in a way that preserved validity and reliability. Secondly, as factor analytic techniques revealed that specific chatbot features tended to cluster under each latent factor and focus group discussions supported the unique contributions of each feature to the chatbot interaction experience, item selection was also performed with the intent of ensuring the representation of each chatbot feature in the dedicated factor with at least one item.

In this study, item evaluation and selection was performed iteratively on the basis of *alpha if the item is dropped* (reliability), *corrected item-total correlations* (item-scale correlations) and item means and variances. First, items which resulted in a significant increase in overall alpha were dropped in order to increase the reliability of each factor. If this step was not able to sufficiently shorten scale length, items with item-total correlations below 0.5 were dropped in order to exclude items that did not correlate adequately with the other items in the factor. If this step was also unsuccessful at identifying ‘bad’ items, items with item means

closer to the midpoint of the 5-point Likert scale exhibiting significant variance were retained. Using this method, 21 items were dropped, resulting in the refinement of factors 1 (communication quality) and 2 (response quality).

For factor 3, or perceived privacy, only one of three items was retained. While single-item measures are controversial in that they may not exhibit sufficient validity, sensitivity and reliability, there is evidence in the literature to support the use of single-item measures if the construct is narrowly defined (Bergkvist & Rossiter, 2007; Drolet & Morrison, 2001). Perceived privacy was defined as 'the extent to which the user feels the chatbot protects one's privacy' (Table 7). While perceived privacy may be informed by many different aspects of the interaction, the items tapping this construct specifically ask the user if they feel that their privacy is being protected when interacting with the chatbot and thus remains narrowly-defined and concrete in this regard. Additionally, the item retained for this factor exhibits a well-balanced distribution of responses with a central tendency close to the scale midpoint and significant variance, suggesting that it may be a sensitive enough measure on its own. The single item retained for this factor thus seems adequate to assess the factor perceived privacy.

For factor 4, or perceived speed, only one of three items was retained as was done for factor 3. Perceived speed was defined as 'the ability of the chatbot to respond timely to user's requests' and thus targets specifically the subjective assessment of how quickly the chatbot responds to user input. While the item retained for this factor exhibits a relatively high central tendency and is positively skewed, this is to be expected as most chatbots have been designed to respond quickly. However, the presence of variance implies that respondents still differ in their evaluations of perceived speed. The single item retained for this factor thus seems adequate to assess the factor perceived speed.

The above procedure resulted in a 17-item questionnaire (Table 16) comprised of four factors: communication quality ($\alpha = 0.73$), interaction quality ($\alpha = 0.91$), perceived privacy and perceived speed. Reliability could not be computed for perceived privacy and perceived speed as they are single-item scales.

Table 16*: Preliminary 17-item questionnaire to assess USIC*

Factor	**Item (Y)**	**Item**
Communication quality	1	It was clear how to start a conversation with the chatbot.
	2	It was easy for me to understand how to start the interaction with the chatbot.
	4	The chatbot was easy to access.
	5	The chatbot function was easily detectable.
	10	I had to rephrase my input multiple times for the chatbot to be able to help me (R)
	11	I had to pay special attention regarding my phrasing when communicating with the chatbot (R)
Response quality	7	Communicating with the chatbot was clear.
	15	The chatbot maintained relevant conversation
	18	The chatbot was able to make references to the website or service when appropriate
	24	I find that the chatbot understands what I want and helps me achieve my goal
	25	The chatbot gave me relevant information during the whole conversation
	30	The chatbot only gives me the information I need
	33	When the chatbot encountered a problem, it responded appropriately
	34	I found the chatbot's responses clear
	37	I feel like the chatbot's responses were accurate
Perceived privacy	21	I believe this chatbot maintains my privacy
Perceived speed	41	The chatbot is quick to respond

4.4. Discussion

This thesis sought to work towards the development of a tool that is able to assess overall user satisfaction with information chatbots as well as indicate with more specificity which aspects of the chatbot interaction users are satisfied with (or not). The purpose of this study was therefore to evaluate the current version of the questionnaire by identifying the underlying factor structure and providing preliminary evidence in support of its validity and reliability.

4.4.1. Interpreting the factor structure.

Factor analysis revealed a structure comprising four factors that best captures the data acquired from participants. The factors were interpreted as follows: **(1) communication quality**, or the ease with which the user can initiate an interaction with the chatbot and communicate one's request, **(2) response quality**, or the quality of the response provided by the chatbot after the user has provided some form of input, **(3) perceived privacy**, or the extent to which the user feels that their privacy is being protected during the interaction and **(4) perceived speed**, or how quickly the chatbot seems to respond to a given input. The four-factor structure obtained largely resembles the five-factor structure proposed in Table 9. As the model in Table 9 was devised on the basis of focus group discussions and relevant literature, the overlap suggests a degree of construct validity for the obtained four-factor structure.

As described earlier, the analyses were conducted using a person × item dataset instead of a design × item dataset and this has significant implications for how the results are interpreted. While classic psychometric theory does indeed make use of person × item data to assess differences between individuals, evaluations in the field of HCI and human factors are performed not on people but on designs and may be more appropriately referred to as *design-metrics*. The goal of this study was to develop a questionnaire that assesses differences between

chatbot designs along user satisfaction, therefore the object of evaluation is the design itself. Unlike psychometrics which utilise person × item matrices, design-metrics require design × item dataset thus the analysis should have been performed on a design × item dataset. However, producing such a dataset requires a substantially larger number of chatbots to be tested by participants but this was not feasible given the practical constraints faced in the current study. Given this distinction, the structure obtained in this study describes individual differences between participants regarding the manner in which they evaluate different aspects of chatbot interactions.

While the factor structure tells us about how people differ in how they subjectively evaluate chatbots, it does not inform us about how the chatbot designs themselves differ based on the subjective evaluations provided by participants. This distinction is nuanced albeit important when interpreting the results obtained in this study. That being said, the obtained structure is not entirely without value. The four factor-structure acquired via a psychometric approach tells us that the individuals likely differ along these four different dimensions when evaluating interactions with information chatbots. Such a structure can be informative in certain contexts, such as when comparing different user groups in order to identify which dimension is associated with the most inter-individual variance in user satisfaction scores. Additionally, while this does not imply that information-chatbots themselves differ along these same four dimensions, it is possible that the same structure may be obtained if a design × item dataset is analysed, although this is subject to further study.

4.4.2. Limitations.

High latent factor correlations. It was found that the four latent factors obtained were highly correlated, which indicates a lack of discriminant validity between the individual sub-scales. This is a significant issue as the intent was to produce a diagnostic measure - the high correlations between latent factors imply that based on the current factor structure obtained, the questionnaire is unable to distinguish between evaluations associated with different aspects of the interaction. Results obtained from the focus groups support the notion that while many chatbot features are closely linked with one another, they are still perceived to be qualitatively distinct. Additionally, these chatbot features have been clustered in meaningful ways as can be observed through the structure obtained though exploratory factor analysis and thus do appear to be assessing distinct dimensions of the chatbot interaction. One interpretation of this finding is that while participants generally differed along whether they were satisfied with the chatbot or not, only to a small degree were participants able to indicate specifically what they were satisfied or not with. This is, however, merely one possible explanation. Several methodological limitations associated with the present study may have been responsible for the high correlations obtained among latent factors, which are discussed later in this section.

Lack of convergence during confirmatory factor analysis. An interesting observation was that the model proposed in Table 9 did not converge during confirmatory factor analysis despite its high overlap with the structure obtained during exploratory factor analysis. One reason could be due to the minor differences found between the two structures which suggests that the hypothesised model may have been inaccurate. While communication quality was expected to also reflect the feature expectation setting, this feature was instead captured under response quality. Additionally, the feature graceful breakdown was expected to be captured as a separate factor but instead was placed under the factor response quality. However, it may also be the case that the specified model was indeed accurate while the limitations of the present

study may not have allowed the model to converge successfully. The methodological limitations of the present study are discussed below.

Methodological limitations.

Sample size. The sample size in this study may not have been sufficient to allow the model to converge and reveal the expected distinct latent factors underlying the questionnaire. It is therefore recommended that studies following up these results be conducted with a substantially larger sample size.

Item quality. The way in which participants responded to the questionnaire may have been responsible for the lack of discriminant validity found in the final structure. One factor that could have influenced this was item quality. Due to time constraints, the item pool was not refined on the basis of the feedback obtained during the item review conducted during the series of focus groups. While the feedback on the items were generally positive, there were some comments that could have been taken into consideration in order to further improve the item pool. As such, item quality may have been compromised and this could have had an impact on the way participants responded to the questionnaire in this study. It would be prudent for future studies to first refine the item pool and consider generating more items in order to improve item quality.

Usability testing paradigm. Another factor contributing to the manner in which participants responded to the questionnaire relates to the usability testing procedure used in this study, which was found to be lacking in two ways. First, each participant was only required to perform one information-retrieval task for each chatbot in the interest of time. However, it was observed that participants did not interact with the chatbot for a sufficient period of time and/or did not encounter certain situations, which made it difficult for them to evaluate certain aspects of the interaction and thus to respond to those items. A noteworthy example of this was related

to the chatbot feature graceful breakdown - many participants did not encounter a situation in which the chatbot "broke down", which led to a variety of different responses from participants. Using the Likert scale provided, some participants used the midpoint to indicate neutrality (as they did not encounter this situation), while others used the extreme negative ("1") to indicate that it did not respond this way or the extreme positive ("5") to indicate that satisfaction as they did not encounter difficulties at all. One task is likely not indicative of the chatbot's performance to the respondent and this may have influenced the way in which participants responded to the items. The second issue was far more specific and relates to one particular chatbot feature - accessibility. During the experiment, participants were given a hyperlink to direct them to the relevant chatbot. This did not allow them to adequately assess whether or not the chatbot was easily accessible, especially if the chatbot was on Facebook as opposed to a website. It is therefore recommended to devise a relatively comprehensive set of information-retrieval tasks for each chatbot to allow respondents to gain a clearer impression of the chatbot such that their responses accurately reflect their subjective assessment of the interaction. If the acquired data did not accurately capture the participants' evaluations of the chatbot interactions, this explains why the model would not converge even if it was correct.

Approach to dataset. Finally, it may be the case that the pattern of latent factor correlations obtained via the psychometric approach taken in this study may not be found when using a design-metrics approach. It may be the case that while individuals do not greatly differ in the manner in which they evaluate different aspects of the chatbot interaction, chatbot designs *do* differ significantly in the subjective evaluations made regarding each aspect of the chatbot interaction, thus potentially resulting in more moderate correlations and thus higher discriminant validity of individual sub-scales. Subsequent study is required using a design-metrics approach whereby scores are analysed across chatbots rather than participants in order to obtain a more valid correlation matrix between latent factors. The nature of the dataset used

for analysis may also explain why the hypothesised model did not converge. The models hypothesised in Tables 8 and 9 were devised with the intent of capturing differences between information-chatbot designs (design-metrics) rather than differences between individuals in the way they evaluate chatbot designs (psychometrics). As the dataset was prepared consistent with a psychometric approach, the way in which the data was analysed may not have allowed for the appropriate testing of the hypothesised models. It may still be the case that the hypothesised model converges through confirmatory factor analysis if the appropriate design × item dataset is used for data analysis in further study.

4.4.3. Future directions.

As the results obtained in this study are preliminary in nature, several actions need to be taken up in subsequent studies in order to arrive at a valid and reliable measure of user satisfaction that is able to distinguish between different information-chatbots as well as indicate satisfaction associated with specific aspects of the chatbot interaction.

First, it is necessary to pay additional attention to the refinement of the usability testing paradigm in which the questionnaire will be administered as this can have a significant impact on the quality of answers provided by respondents and by extension, on the quality of data upon which analyses are performed. Addressing the relevant methodological limitations discussed in the previous section will allow us to acquire responses of higher quality from participants.

Importantly, the next study should perform all analyses on a design × item dataset. In order to obtain enough observations, it is important to test a substantially larger number of chatbots than was utilised in the present study. The dataset is then prepared by collapsing responses to each questionnaire item across participants for each tested chatbot. It is not nearly as important to increase the number of participants as it is to increase the number of chatbots

tested as the desired outcome is design-metric and not psychometric in nature. The next study should also seek to confirm the underlying structure of the questionnaire from a design-metric perspective. In addition to testing the four-factor structure obtained through exploratory factor analysis, subsequent analyses should also test the models specified in Tables 8 and 9 through confirmatory factor analysis and compute fit indices. It is expected that the factors should correlate not highly but moderately in order to exhibit adequate discriminant validity.

As the development of this questionnaire is still in its primary stages, it still requires further evidence in support of its validity and reliability before it can be implemented. The questionnaire can be administered to respondents alongside alternative standardised usability measures that have already been shown to have high validity and reliability in order to support the questionnaire's criterion validity. A good candidate for this is the System Usability Scale (SUS) (Brooke, 2013). It would be expected to correlate moderately with the questionnaire in development because it is a valid and reliable but non-diagnostic measure of subjective usability assessment. Regarding reliability, efforts should be made to replicate the internal consistency of these factors to ensure stability of the reliability coefficients across different samples. Of importance is assessing the reliability of the single-item factor through test-retest reliability as traditional reliability assessments are not applicable.

5. General Discussion

This thesis set out to continue the work initiated by Tariverdiyeva and Borsci (2019) with the intent of developing a preliminary version of a tool that can be used to assess user satisfaction with information chatbots.

First, a pre-experimental phase and a series of focus groups were undertaken to directly follow up the qualitative literature review conducted by Tariverdiyeva and Borsci (2019) in order to further refine the list of chatbot features that are important for shaping satisfaction with information-chatbot interactions and establish content adequacy of the list of chatbot features and preliminary item pool. Through these activities, the list of chatbot features pertinent to the USIC construct was refined considerably. Upon analysing the chatbot features that were retained, it was concluded that the primary motivation that potential end-users have for interacting with information chatbots is productivity, and therefore it is expected that information chatbots behave in a manner that is both highly effective and efficient while addressing users' information-retrieval goals.

However, this does not imply that all excluded features are rendered unimportant – in fact, some of these features still need to be taken into consideration by chatbot developers in order to provide the best experience for end-users. Some features, such as privacy & security, information accuracy, meeting of neuro-diverse needs and ethical decision-making, are difficult to evaluate as an individual end-user but are still essential to a good chatbot. Other features, such as a minimally appropriate language style and response time, are often taken as a given and need not be evaluated through a questionnaire although their inclusion in a chatbot is crucial. Therefore, in addition to the retained chatbot features, several excluded features such as the ones mentioned above can be used to inform a checklist (e.g. Ferman, 2018) for designers at different stages of customer service chatbot development to ensure user expectations are met.

Of note was the resounding confirmation that personality and enjoyment were considered unequivocally unimportant for information-chatbots from the perspective of potential end-users despite significant evidence of the contrary found in the literature. This can be attributed to the fact that the primary motivation that users have for engaging with such chatbots is productivity. These findings suggest that even among chatbots, the factors relevant to determining user satisfaction may critically depend on the specific type of chatbot. Paikari and van der Hoek (2018), in addition to information, identified two other types of chatbots based on their function: collaboration and automation chatbots. Fadhil (2018) instead classified chatbots based on the domain they were designed for, ranging from e-commerce and productivity to health and entertainment and identified domain-specific patterns and differences. Folstad, Skjuve and Brandtzaeg (2019) suggest that chatbots also differ in terms of how long the relation lasts between the chatbot and its user i.e. short and long. Such differences between chatbots result in significant and meaningful variations in overall context of use, resulting in different weights assigned to the relative importance of different chatbot features for a satisfactory interaction. Future studies can investigate the use and potential development of different questionnaires in order to evaluate user satisfaction with other types of chatbots and even other conversational interfaces such as digital personal assistants.

The present findings confirm the notion that factors that shape user satisfaction do indeed differ between traditional graphical ‘point-and-click’ user interfaces such as websites and smartphone applications, and natural language interfaces. While the factors pertinent to website usability include constructs such as readability and navigability (Lee & Kozar, 2012), it appears that natural language interfaces differ in fundamental ways, changing the nature of the interaction in such a way that the factors that determine the perceived usability of conversational interfaces such as chatbots are, in fact, different and reinforces the need for a dedicated measurement tool such as the one developed in this study.

After refining the list of chatbot features and generating an item pool, this thesis also set out to conduct a preliminary evaluation of the questionnaire. The item pool was administered as a post-test questionnaire and factor analytic techniques were applied to the acquired data, revealing a four-factor solution that best captured the participants' responses to the questionnaire: *communication quality*, *response quality*, *perceived privacy* and *perceived speed*. However, the psychometric approach taken in this study means that the resulting factor structure represents how individuals differ in how they evaluate interactions with information chatbots. While valuable in its own right, the desired outcome hinges upon the adoption of a design-metric approach and a structure that instead reflects how user satisfaction differs between different information chatbot designs, the achievement of which lies in the work of future studies. Follow-up studies can be conducted in line with the discussion of the present study's limitations and recommendations presented in the previous section in order to arrive at the desired tool. It is expected that the resulting questionnaire can be implemented as a diagnostic measure of user satisfaction with information chatbots. It will be able to provide not only an overall score of user satisfaction but can reveal which aspects of the interaction led to the user's (dis)satisfaction. Chatbot technology is its in infancy stages and studies have only recently begun to take a user-centred approach to chatbot research on account of the fact that existing chatbots often fail to impress their users. In this context, such a measure would be highly beneficial for chatbot developers who wish to target and improve various aspects of the user experience associated with information chatbots such as customer service bots.

6. Conclusion

While the current version of the questionnaire requires further validation efforts before it can be implemented, this thesis has contributed significantly to the development of a dedicated, diagnostic and standardised measure that can assess user satisfaction associated with the myriad of information chatbots that are rapidly gaining momentum in the domain of customer service. The findings not only support the fundamentally distinct nature of natural-language interfaces but also align with previous studies that have explored user perceptions and expectations of chatbots in revealing that the key driver of adoption for information chatbots is, in fact, productivity. Chatbot developers should keep in mind that the entire user experience revolves around creating an effective and efficient way for the user to achieve their goal in a way that is also superior to the technologies that we are all too familiar with.

References

Abbas, A. (2019, February 12). Retrieved from https://chatbotslife.com/chatbot-2019-trends-and-stats-with-insider-reports-fb71697deee4

AbuShawar, B., & Atwell, E. (2016). Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, *19*(2), 373-383.

Assink, L. M. (2019). *Exploring Users' Perception of Chatbots in a Mobile Commerce Environment: Creating a Better User Experience by Implementing Anthropomorphic Visual and Linguistic Chatbot Features* (Bachelor's thesis, University of Twente).

Beaver, L. (2017, May 11). Chatbots are gaining traction. Retrieved from https://www.businessinsider.com/chatbots-are-gaining-traction-2017-5?international=true&r=US&IR=T

Beaver, L. (2017, May 11). Chatbots are gaining traction. *Business Insider*. Retrieved from https://www.businessinsider.com/chatbots-are-gaining-traction-2017-5?international=true&r=US&IR=T

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. Journal of marketing research, 44(2), 175-184.

Beriault-Poirier, A., Tep, S. P., & Sénécal, S. (2018, October). Putting Chatbots to the Test: Does the User Experience Score Higher with Chatbots Than Websites?. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications* (pp. 204-212). Springer, Cham.

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.

Bosley, J. J. (2013). Creating a short usability metric for user experience (UMUX) scale. Interacting with Computers, 25(4), 317-319.

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

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: changing user needs and motivations. *interactions*, *25*(5), 38-43.

Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

Brown, I., & Jayakody, R. (2008). B2C e-commerce success: A test and validation of a revised conceptual model. *The Electronic Journal Information Systems Evaluation*, *11*(3), 167-184.

Chopra, S., & Chivukula, S. (2017, September). My phone assistant should know I am an Indian: influencing factors for adoption of assistive agents. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (p. 94). ACM.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. Journal of econometrics, 183(1), 31-57.

de Haan, H., Snijder, J., van Nimwegen, C., & Beun, R. J. (2018). Chatbot Personality and Customer Satisfaction.

DeVellis, Robert F. Scale development: Theory and applications. Vol. 26. Sage publications, 2016.

Diederich, S., Brendel, A. B., Lichtenberg, S., & Kolbe, L. (2019). DESIGN FOR FAST REQUEST FULFILLMENT OR NATURAL INTERACTION? INSIGHTS FROM AN EXPERIMENT WITH A CONVERSATIONAL AGENT.

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research?. Journal of service research, 3(3), 196-204.

Fadhil, A. (2018). Domain specific design patterns: Designing for conversational user interfaces. arXiv preprint arXiv:1802.09055.

Fadhil, A., & Schiavo, G. (2019). Designing for Health Chatbots. *arXiv preprint arXiv:1902.09022*.

Ferman, M. (2018). Towards Best Practices for Chatbots.

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

Følstad, A., & Skjuve, M. (2018). Business and pleasure? Relational interaction in conversational UX.

Følstad, A., & Skjuve, M. (2019, August). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*(p. 1). ACM.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018, October). What makes users trust a chatbot for customer service? An exploratory interview study. In International Conference on Internet Science (pp. 194-208). Springer, Cham.

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2018, October). Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design. In *International Conference on Internet Science* (pp. 145-156). Springer, Cham.

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2018, October). Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design. In International Conference on Internet Science (pp. 145-156). Springer, Cham.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304-316.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational research methods, 7(2), 191-205.

Hendriks, F. The effect of chatbot introduction on user satisfaction.

Jain, M., Kota, R., Kumar, P., & Patel, S. N. (2018, April). Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 468). ACM.

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

Koufteros, X., Babbar, S., & Kaighobadi, M. (2009). A paradigm for examining second-order factor models employing structural equation modeling. International Journal of Production Economics, 120(2), 633-652.

Kuligowska, K. (2015). Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. Professionals Center for Business Research, 2.

Lee, Y., & Kozar, K. A. (2012). Understanding of website usability: Specifying and measuring constructs and their relationships. Decision support systems, 52(2), 450-463.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction, 14(3-4), 463-488.

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. International Journal of Human–Computer Interaction, 34(12), 1148-1156.

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

Liao, Q. V., Hussain, M. U., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., ... & Geyer, W. (2018, April). All Work and No Play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 3). ACM.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199-204.

Luger, E., & Sellen, A. (2016, May). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286-5297). ACM.

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. arXiv preprint arXiv:1511.05604.

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, *20*(6), e10148.

Muresan, A., & Pohl, H. (2019, April). Chats with Bots: Balancing Imitation and Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (p. LBW0252). ACM.

Nijholt, A., Niculescu, A. I., Alessandro, V., & Banchs, R. E. (2017). Humor in human-computer interaction: a short survey.

Ogawa, R. T., & Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. Review of educational research, 61(3), 265-286.

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

Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, 89-97.

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

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

Revelle, W. R. (2017). psych: Procedures for personality and psychological research.

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., ... & Jinks, C. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & quantity, 52(4), 1893-1907.

Sheehan, B. T. (2018). *Customer service chatbots: Anthropomorphism, adoption and word of mouth* (Doctoral dissertation, Queensland University of Technology).

Skjuve, M., & Brandzaeg, P. B. (2018, October). Measuring User Experience in Chatbots: An Approach to Interpersonal Communication Competence. In *International Conference on Internet Science* (pp. 113-120). Springer, Cham.

Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. B. (2019). Help! Is my Chatbot Falling into the Uncanny Valley?: An Empirical Study of User Experience in Human-Chatbot Interaction. *Human Technology*, *15*(1).

Smestad, T. L. (2018). *Personality Matters! Improving The User Experience of Chatbot Interfaces-Personality provides a stable pattern to guide the design and behaviour of conversational agents* (Master's thesis, NTNU).

Smestad, T. L., & Volden, F. (2018, October). Chatbot Personalities Matters. In *International Conference on Internet Science* (pp. 170-181). Springer, Cham.

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

Sörensen, I. (2017). Expectations on chatbots among novice users during the onboarding process.

Steinbauer, F., Kern, R., & Kröll, M. (2019, July). Chatbots Assisting German Business Management Applications. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 717-729). Springer, Cham.

Suthar, S. (2019, April 22). Retrieved from https://acquire.io/blog/chatbots-trends/

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

Trivedi, J. (2019). Examining the Customer Experience of Using Banking Chatbots and Its Impact on Brand Love: The Moderating Role of Perceived Risk. *Journal of Internet Commerce*, *18*(1), 91-111.

Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. Educational and Psychological Measurement, 58(4), 541-568.

Verney, V., & Poulain, A. (2018). *Building brand resonance with chatbots: assessing the importance of giving your bot a human personality* (Master's thesis, Handelshøyskolen BI).

Walker, M. A., Passonneau, R., & Boland, J. E. (2001, July). Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 515-522). Association for Computational Linguistics.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806-838.

Yang, X., Aurisicchio, M., & Baxter, W. (2019, April). Understanding Affective Experiences With Conversational Agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 542). ACM.

Zamora, J. (2017, October). I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction*(pp. 253-260). ACM.

Facebook. (July 24, 2019). Number of monthly active Facebook users worldwide as of 2nd quarter 2019 (in millions) [Graph]. In *Statista*. Retrieved September 19, 2019, from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Appendix 1: Focus Groups

1.1 Demographic questionnaire

Session: _____ Participant ID: _____

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

Age _____

Gender M / F

Nationality _____

Field of study _____

Have you used a chatbot before? Yes / No

If yes, then answer the two questions below.

How often do you use chatbots?

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

How would you rate your previous experiences with chatbots?

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

1.2. Participant Information Sheet

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

Principal investigator: Divyaa Balaji

Co-investigator: Dr Simone Borsci

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

Purpose of the study

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

Your role as participant

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

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

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

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

Personal data

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

Videotaping and Questionnaire

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

All data will be made anonymous before stored and secured on a separate hard drive to which the research team and supervisor will have access during the research period while writing bachelor and master theses. When data evaluation is finished, the access will belong solely to the supervisor. The research has the potential to be published and therefore, the data will have a retention period of approximately 12 months, when it is expected to be published. During the retention period, only the supervisor will have access to it.

Ethical review of the study

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

Contact details

Principal Researcher	*Co-Investigator*
Divyaa Balaji	Dr. Simone Borsci
d.balaji@student.utwente.nl	s.borsci@utwente.nl

Consent Form for Assessing user satisfaction with chatbot interactions

YOU WILL BE GIVEN A COPY OF THIS INFORMED CONSENT FORM

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

Signatures

_____ _____ _____

Name of participant [printed] Signature Date

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

_____ _____ _____

Researcher name [printed] Signature Date

1.3. List of chatbot features (n = 21)

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

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

1.4. Preliminary item pool

No.	Item	Factor(s)	Comments
1	The time of a response was reasonable.		
2	The chatbot solved my problems instantly.		
3	I felt that I could trust the chatbot.		
4	The interaction with the chatbot felt secure in terms of privacy.		
5	I feel like the chatbot's responses were accurate.		
6	I found the chatbot's responses clear.		
7	The chatbot gave relevant information during the whole conversation		
8	The style of language used by the chatbot felt appropriate.		
9	The interaction with the chatbot felt like an ongoing conversation.		
10	The amount of received information was neither too much nor too less.		
11	The interaction with the chatbot felt easy.		
12	I had to rephrase my input multiple times for the chatbot to be able to help me.		
13	The chatbot was easy to spot on the website.		
14	It was clear how to start a conversation with a chatbot.		
15	Communicating with the chatbot was clear.		
16	The chatbot guided me to the relevant service.		
17	I was adequately updated about my task progress.		

18	I felt that my intentions were understood by the chatbot.		
19	The chatbot could handle situations in which the line of conversation was not clear		
20	The chatbot seemed like a human with its own personality		
21	I enjoyed interacting with the chatbot		
22	My waiting time for a response from the chatbot is short.		
23	The chatbot is able to answer any questions within a few seconds.		
24	The chatbot reassures me that I can trust this technology.		
25	I believe the chatbot is informing me of any possible privacy issues		
26	I believe that the chatbot only states reliable information.		
27	The chatbot only states understandable answers		
28	The chatbot is good at providing me with a helpful response any point of the process.		
29	The chatbot is answering with the right amount of formality		
30	The chatbot was able to keep track of context.		
31	The chatbot gives me the appropriate amount of information.		
32	I had to put in only minimal effort to use the chatbot.		
33	I had to pay special attention regarding my phrasing when communicating with the chatbot.		
34	The chatbot function is easily detectable for the user on the website		
35	The design of the chatbot guided me into starting a conversation		
36	I was immediately aware of what information the chatbot can give me.		

37	The chatbot is using hyperlinks to guide me to my goal		
38	The chatbot is giving me feedback about the status of my request		
39	The chatbot was able to guide me towards my goal.		
40	The chatbot explained gracefully that it could not help me		
41	The chatbot communicated in a pleasant way with me		
42	The chatbot made it fun to research the information		
43	The chatbot is quick to respond.		
44	I trust this chatbot.		
45	I believe that this chatbot maintains my privacy.		
46	I feel like the chatbot's responses were accurate.		
47	The chatbot's responses were easy to understand.		
48	The chatbot provided relevant information as and when I needed it.		
49	The chatbot communicates with an appropriate language style.		
50	The chatbot maintains relevant conversation.		
51	The chatbot only gives me the information I need.		
52	I find the chatbot easy to use.		
53	It is easy to tell the chatbot what I would like it to do.		
54	It is easy to find the chatbot on the website.		
55	I find it easy to start a conversation with the chatbot.		

56	It is clear to me what the chatbot can do.		
57	The chatbot keeps me aware of what it is doing.		
58	I find that the chatbot understands what I want and helps me achieve my goal.		
59	When the chatbot encountered a problem, it responded appropriately.		
60	I found the chatbot to be likeable.		
61	The chatbot was fun to interact with.		

1.5. Session script

[Introduction] Hello everyone! Thank you for coming here today.

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

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

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

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

[Demographics] Before we jump into the discussion, please fill out this short form for us about yourselves.

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

First, everyone's opinion is valued and important for this topic. There is also no such thing as a right or wrong opinion. Second, everyone should get the chance to talk without interruptions. Third, this is a discussion and thus, you do not have to talk to me the whole time. It is perfectly fine to look and talk to each other directly.

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

[Introduction] A chatbot is a kind of software program running on artificial intelligence. They're expected to be able to simulate a human-like conversation, using natural language. Chatbots generally return a response based on either voice or text input from a user.

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

<< Have any of you used chatbots before? >>

[Interactive demonstration] We're going to spend about 10 minutes testing two of these chatbots right now. If you haven't used one before, this is your chance to get familiar with them. If you have, then you can refresh your memory about them. So I'd like you to discuss

and agree on what to ask the chatbot, essentially decide how to interact with it, and I will communicate with the chatbot.

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

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

<< Any questions or doubts about chatbots? >>

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

(Give each individual the list of 11 factors)

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

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

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

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

(Repeat with the remaining 11 factors)

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

(Give each individual the list of items)

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

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

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

<< Are there any questions? >>

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

1.6. Transcribed document for all focus groups

Focus Group 1

Note: no one has used a chatbot before – all input based on first impression with Finnair's chatbot as a reference

FG1 - interaction with Finnair's chatbot:

- Output: chatbot returned an almost entirely irrelevant answer to our request
- Noticed that chatbots may require specific information and that's the reason it did not give the user the information he or she was looking for – users feeling the need to rephrase and be more careful about how the request is put across; this user gives the impression that they feel responsible for making themselves understood
- Pointed out that the chatbot mentioned early on that it wanted 'simple' and 'direct' questions but the question asked seemed to meet that criteria and yet it was unable to provide a helpful response – confusion about what the instructions meant; the instructions at the beginning were likely meant to set expectations so the user could modify his or her mental model accordingly but for this particular chatbot, despite doing so, the interaction did not meet the set expectations which created confusion in the user – important to set expectations but also for the chatbot to act consistently with explicitly set expectations
    - Another user also agreed that the information given to the chatbot should have been enough to get a decent response – other users believe that the chatbot should have understood their phrasing even if it wasn't perfect
- **Accuracy** – users expected accurate information from the chatbot too
- What they liked about it: **fast**, appreciated that the users were told in the beginning to direct more complicated questions to its 'human agents' (expectation setting?)
- What they didn't like: not helpful at all (it gave a response but the response was not at all relevant and provided no useful information), might just go to the website directly instead (perhaps it's important to evaluate whether the user would prefer using the relevant service platform over the chatbot)

FG1 - factor list:

- Unclear
    - Ability of the chatbot to gracefully handle unexpected input – what does this mean?
    - Ability of the chatbot to solve problems on the spot – how would a chatbot be able to do this anyway without human help? What kind of problems? Perhaps it's not relevant because "according to me, I think a chatbot is supposed to help you solve a problem and not actually solving it" – this user has a specific expectation for a chatbot in that she does not expect it to be able to solve her problems and merely provide her with the information; based on our interaction with Finn, she has gathered that a chatbot cannot solve our problems directly, and it can't do what a human being can.
    - Process tracking – this user interpreted this factor as perhaps updating the user if there are any changes to be aware of (which would be useful) but this is not how the factor was meant to be interpreted. "It would get annoying if the bot took a long time to search for information and also keep telling me which status it is in now…that's not helpful at all" – another user thinks it might be helpful but that the user should have a choice about whether they want to use it or not; other user agrees that if it's customizable, then it could be useful and provide an advantage over traditional websites because "users need some kind of incentive to actually use a chatbot."
- Unimportant

    - Personality – why? For specialised chatbots such as these, the user has a task-oriented goal in mind and is also aware that it is indeed merely a bot, he just expects the chatbot to help him achieve his goal and no more; another user also thinks so – he thinks "it's not right". However, another user thinks that in some domains, a chatbot with an appropriate personality might be helpful in helping users trust and engage with the chatbot more e.g. mental health and the fourth user also disagrees but agrees that it could be useful in certain situations.
    - Enjoyment – refer to above. It is not necessary nor is it expected but it could be "nice to have" although it could come in handy in some situations. The goal is to "get some quick answers…so it's not a top priority for me." Another user disagrees – "if I enjoy the interaction, not in a fun, funny kind of way but that it was easy and just nice to use, then I would be more likely to use it again" – the other user also agrees with this user. Note: both of the users that value enjoyment are female while the users that do not place much importance on this factor are male – just an observation, I'm not being sexist.
- Important
    - Ease of use, in general – user points out that it was helpful that the chatbot gave instructions for its use right at the beginning and essentially guiding the user on how to interact with it and ask questions (is this expectation setting though?) and "if it's not easy to use, then I would never use it again."
    - Factors 3 to 5 (trust, privacy and security and perceived credibility) – one user pointed these out as very important.
    - Customizability – users seem to want to be able to choose how the chatbot interacts with you. Also the option of explicitly saving and remembering preferences or settings, especially for chatbots on commonly-visited websites. Not the same as learning from the user – not comfortable with this. Having control over what the bot can and can't do (setting limits and controlling what the bot knows) and being aware of what the bot does is important because otherwise it's scary. Another user thinks it shouldn't remember his settings at all – he wants to be "free" when using it – choice is important.
- Factor overlap
    - Ease of use, understandability, ease of starting a conversation, flexibility of linguistic input
    - Privacy and security, trust
    - Response time, on the fly problem-solving

FG1 – items

- Use of words such as "reasonable" and "appropriately" – vague, subject to interpretation; try to be more specific?
    - "responded appropriately" but what is an appropriate response in this case? User was not sure how to interpret this item – no one else had this issue with this item; also couldn't match it to a factor.
- "provided relevant information as and when I need it" – confusion; user expects to ask a chatbot for information and receive a response immediately; misinterpretation and needs to be reworded.
- "the chatbot reassures me that I can trust this technology" – "a bit creepy – I trust something because it seems trustworthy, not because it tells me that it is trustworthy" – rephrasing necessary.
- Several items were matched to more than one factor
    - Item 1 matched to factors 1 and 2

- 'Chatbot's responses were clear' matched to factors 6 and 8 – others only matched it to 6
- 'I feel like I can trust this chatbot' matched to factors 3, 4 and 5 (all trust-related factors) – "trust can be interpreted in many ways…trust is pretty broad" – no one else had this issue but consistent with our expectations; after explaining our rationale, all users believe it is important to retain the distinctive trust factors and "you could even get rid of the general trust factor"
- When asked about ease of use – all users agree that ease of use is comprised of different aspects, like it can be easy to use in one aspect but not in another so more questions are needed to measure this adequately.

Focus group 2

Note: one user is in the midst of developing her own chatbot (E); another user is heavily biased towards Apple's Siri (S)

FG2 - interaction with Finnair's chatbot:

- User E didn't expect the chatbot to introduce himself and the instructions it provided on how to use it because she's not used to it – most chatbots don't do it. Thinks it could be helpful when the chatbot clarifies itself on what it's about and what it can do because people can be confused about what to expect and do.
- User pointed out that despite following the instructions about the kinds of questions it can answer, the chatbot still got it wrong – agreement from other users. Other users point out the need to be pretty specific about the thing you're asking if you want a decent response, which can be inconvenient.
- User S would not book a flight using Finnair's chatbot - she says she'll need too much time to think about how to phrase the question to the chatbot in the right way and she usually doesn't have that kind of time or patience because the main reason she uses Siri is that it saves her time. User E chimes in by saying that you need to think about all the information you're including in the question so that it can understand you.
- Importantly, it has to be easier than finding the information yourself (through whatever alternatives exist)
- They liked Finn – he had a "good attitude" and was "sweet" while "people from the airport aren't that nice to you"
- Users agreed that there was too much information being presented too quickly and that this might be a problem, especially for the older population who might not be able to keep up. They seemed to find the information overload both surprising and overwhelming.

FG2 - factor list:

- Maxim of relation - what do you mean by "at each stage"? Is it about being systematic? User E says it's similar to perceived credibility (?)
- Unimportant
    - Factor 3 – user S doesn't see the need for trust in these type of chatbots, assuming trust is a more emotional, engaging, personal kind of construct "willingness to engage" ("misinterpretation) – acknowledges that trust is required in the information it's providing
    - Factor 9 – ability to maintain a conversational theme; convenient and smart but not necessary; upon explanation, "needing to keep track of and stay on top of what the user is trying to achieve throughout the conversation – kind of agrees it's important but also works without it. "Keeping track without the need for repetition" as a rephrasing of the factor description – more accurate.

    - Factor 7 – 'relation'? Three out of four people didn't really get it. How is it different from perceived credibility? (User E and another user also agreed) – suggestion to combine factors 5 and 7 into one
    - Factor 10 (amount of information, only information that is relevant, brief, concise being the keywords)? Related to factor 7 (provide the relevant and appropriate contributions - vague)? They don't consider it the same, just related.
    - Process tracking – user S is confused as to why there would be a need for process tracking if the chatbot is supposed to give you an immediate response; another user was also confused about this factor. User E had a different interpretation, thought of the feature as more of an update-giver to keep the user on top of any changes, like a news feed or notification system – otherwise it should just give you direct answers when you ask.
    - Enjoyment – user E doesn't need enjoyment, she just wants information although she acknowledges that this might influence whether someone will continue using it but personally, she doesn't care.
    - Factor 16 – doesn't want it to automatically open tabs or navigate to pages because it's scary; just provide links.
    - Factors 13 and 14 – quite similar? If it's visible, isn't it easy to start the conversation? Needed clarification. Upon clarification, it is important and they're not the same.
- Important
    - Factor 9 – user S hates repeating herself so this is highly important to her.
    - Factors 12 (and 19 (graceful in the face of the unexpected) too – user S thinks these are important too. Users E and S think they're kind of similar.
        - Does 'unexpected input' include typos? If so, that sounds similar to 12 according to them. We need to make the two factors distinct. Highly important that the chatbot understands what you ask even if your input isn't perfect.
    - Factor 5 – the answers should be true and based on fact.
    - User S - understandability is important too. Also giving you information in some sort of logical and systematic order in a way that makes sense to the user, easy to follow.
    - Important that it doesn't give you too much and unnecessary information, with the option of finding more information through links or something.
    - Factor 20 – humanlike; as long as you are made aware that it is a bot and if it is a bot, "I'd rather it not be humanlike, so I know what to do with it" – the key is that the user doesn't want to be tricked.

FG2 – items (refer to audio)

Focus group 3

Note: users likely have had prior experience with a chatbot.

FG3 – interaction with Finnair's chatbot:

- Fast but it might just be faster to go online and look for flights yourself.
- Notes that you need to focus on phrasing based on what the chatbot can and cannot understand.
    - Agreement between users on the above two comments.
- What they liked about it
    - It responds in a nice manner, answers accordingly like "sorry I don't understand that"

- "Felt like you were talking to a human, with the emojis and broken hearts and stuff" – not necessary but nice to have
- When you asked for flights, it gave you not only the destination and timing but also the price – all the relevant information was presented at once so you didn't need to look further.

FG3 - factor list:

- Unclear
  - Trust – what is trust in this (general) context? Vague, subject to interpretation.
- Important
  - Privacy and security
  - Ability to maintain themed conversation – expectations held for AI, keep track of the conversation and not suddenly change or forget
  - Reference to service – basically to get relevant responses, misinterpreted maybe?
  - Visibility – users need to know the chatbot even exists otherwise you wouldn't use it
  - Redirecting to human agent when the chatbot can't handle – this is an example of graceful responses in unexpected situations – connect to reference to service? At this stage, users are aware that chatbots can't handle all their requests but the chatbot should be able to sense when it can't help and act accordingly.
  - Precision, accuracy (???)
- Unimportant
  - Humanlike – is it about personality? As long as the responses make sense, being humanlike in any other sense wouldn't be important because you know it's a bot and you wouldn't expect that of it in this context. Another user says it's nice to have because it's connected to enjoyment – makes for a more enjoyable experience. Unlike call centre or customer service personnel, robots don't have feelings so they can continue to be "nice" and "understanding" and can take your "stupid questions" even when you as the user are frustrated and being intolerable.
  - Engage in on-the-fly problem solving – important that the chatbot keeps you informed what it's currently doing (process tracking); you'd expect the human to solve the problem, not the chatbot. Also the first two factors might be similar. One user said that he'd find it odd if the chatbot responded so quickly because if you'd asked a human that question, it would take at least a bit of time. Another user disagreed because the entire motivation behind using a chatbot is that it has the time advantage over talking to a human. Other user agrees that fast is important.
  - Appropriate language style – should be understandable but doesn't need to use the right "style". Another user disagreed and said that it should be at least kind of appropriate depending on the service that the chatbot represent e.g. funeral home. In the case of Finnair's chatbot, despite the use of emojis, the language was still understandable to most target populations.
    - Adaptive language style? Helps it to be more humanlike. Adds to the personality too. Might be helpful for certain audiences. Personalisation? A more tailored experience, but it does it automatically. Aids the entire experience. Connect to recognition and facilitation of user's goal and intent – only using the information you're giving the chatbot, nothing else to adapt the conversation accordingly.
  - Trust and privacy and security – similar? Also about information accuracy, so in this case also perceived credibility.
  - Context-orientation – ability to maintain themed discussion and understandability – similar? Another users disagrees, saying that they're related but not the same. A lot of the factors overlap in that they're closely related to each other. For example, if

the bot has a humanlike personality, perhaps it's easy to interact with and the interaction is overall enjoyable.

FG3 – items

- "felt secure in terms of privacy" – how can one "feel" secure? A bit vague.
- All items are clear.
- Multiple items generated per factor but one user pointed out that if we're using multiple items to measure one factor, then it's better to reword them so that they don't sound too similar. Could feel a little too repetitive. Rephrasing required.
- Not enough items about visibility (?) – need better items for this factor. Could extend it to other platforms too. Could also talk about placement and accessibility. Factor – general design of the chatbot itself, like how it looks as a pop-up or something e.g. a bubble, moving, etc.
- "felt like an ongoing conversation" matched to ease of use, understandability, personality (human-like), enjoyment, ease of starting a conversation (are any of these factors right?)
- "felt like the chatbot's responses were accurate" matched to several factors too like perceived credibility, themed discussion, etc.
- "believe that the chatbot only states reliable information" matched to perceived credibility, maxim of relation, ability to maintain themed discussion
- Maxim of relation, flexibility of linguistic input, ease of starting conversation, ability to maintain themed discussion -> all of these are associated with human-like behaviour
- "solved problems instantaneously" – response time, on-the-fly problem solving, etc.

Focus group 4

Note: everyone has used a chatbot before.

FG4 – interaction with Finnair chatbot

- Didn't like it
    - Users surprised that it didn't understand the word 'bag' even though it's a commonly used word – it only responded to the specific word 'baggage'; they expect it to be able to understand as many things as possible
    - Another user commented that while this aspect could be improved, the fact that it presented him with options to click on meant that if it didn't understand his text input, he could still get the job done simply by clicking.
    - Didn't like the information overload that happened right at the beginning, would prefer it give information systematically with appropriate time gaps, the user didn't even have time to read. Felt like you were being bombarded.
    - User thinks it doesn't have to be enjoyable, it just has to get the job done. Two others agree. Other user likes that it was sarcastic and funny, makes it enjoyable and pleasant to interact with – it's better to make it a bit more human and fun. Can sometimes be annoying but nice to have. Depends on your mood.
- What did you like?
    - "It was cute"
    - Too fast, too much information.
    - As long as it gets the question, it gets to the point.

FG4 – factor list

- Unclear
    - Response time – time to get the answer or time for the response itself? Need to emphasise its distinction from accuracy and relevance.
- Unimportant

    - Enjoyment – the user is there for information so enjoyment really isn't important here as long as it gets the job done. Might be more relevant for other kinds of chatbots but maybe not for these. Other user really does want enjoyment. For most of the users, enjoyment is a take-it-or-leave-it situation; the task takes priority – as long as it gets what you're saying, doesn't confuse it for something else and what you're searching for is available, it's good.
        - Personality too! Same as above. It's probably more important for the fun/general ones but she still wants it. Her argument is that you have to make chatbots appealing for people to use and one obvious way to do this is to make it more "humanlike". Other user says its relevant but probably depends on the application. Probably has consequences for future use. He also points out that if you're generally in a bad mood or frustrated with the chatbot for not being able to help, making it likeable might help the user tolerate it for longer. Other users seem to agree with this last point.
    - Ease of starting a conversation – redundant because how hard could it be? Another user recounts his experience of how he simply couldn't get a chatbot to understand what he wanted and he would've appreciated some advice at the beginning on how to interact with it, perhaps some options to click on to get started, etc. Need to reword description. Users who have had significant experience with chatbots and also those who are patient and willing to reword will have fewer problems along this aspect – may not be difficult for everyone.
    - Trust, perceived credibility – similar (other users agree); users also agree that aspects of trust should be separate. Holding it responsible, "accountability" (under general trust) – threw off the user. General trust redundant.
    - Understandability and flexibility of linguistic input – similar? After clarification, not the same at all.
- Important
    - Ability to maintain themed discussion – don't want to repeat yourself. Keep context. Remain relevant.
    - Response time
    - Maxim of relation – you want only the relevant stuff. Getting to the point. Information you actually want. You might be interested in more information but it shouldn't bombard you with all of that – pick out the crucial things so just enough information. One user pointed out that if it starts incorporating options and references too much it'll become like an app with a GUI and that's not what he wants from a chatbot, he wants to be able to type and receive actual replies. Some prefer option over typing because it's more efficient, especially if you're in a hurry. Balance is key.
    - Graceful responses to unexpected situations – very humanlike, can handle all kinds of input, even typos – related to flexibility of linguistic input. Maybe reference to service as a graceful response. Should handle breakdowns in its capabilities with grace. Reword this factor. Links would be nice.

GF4 – items

- "communicating with the chatbot was clear" – vague; what can I talk to the bot about or the response itself?
- First two factors and their items – all very similar, "almost the same"; items might be similar but sometimes the factors do need to be addressed separately
- On-the-fly problem solving and graceful responses to unexpected situations – similar?
    - What exactly is "on-the-fly problem solving" in the context of chatbots? Is it about providing responses to your request instantaneously? Is it about solving more

complex problems with ease? It's about providing solutions in the least amount of time possible.
    - Response time may apply to problems but also any other kind of input. Response time – time is the crucial element. On-the-fly problem solving – overcoming obstacles and actually providing a helpful response, hence the connection to unexpected input. *Related to recognition and facilitation of user's goal and intent.*
    - *Not only does it have to solve your problem but also do it in a reasonable time frame (efficiency) – two aspects of problem-solving.*
- All the trust factors could be clustered, one item about privacy matched to all three.
    - Maybe provide a way for users to be sure that divulging personal information is safe and should be done in a certain way. Disclaimers too. Being informed about privacy – full transparency. Also relates to expectation setting (?) – broad?
- Expectation setting – too broad. Anything you should expect from it. Could include privacy, disclaimers, etc.
- Factors 7 (provide relevant and appropriate contributions at each stage) and 17 (process tracking) – confusing? The items (wording) were confusing despite the factors being distinct. Factors could go hand in hand. Process tracking does sound useful, especially if it's going take a while or if it's stuck.
- "solves my problem instantaneously" – five factors.
- "I found the chatbot to be likeable" – 20, 21, **6, 7, 10, 12**; all these factors add up to the pleasantness of the interaction for the user – too broad. What makes the chatbot likeable to interact with, would you use it again? Efficient but not likeable if it satisfied the last four. Likeable as in not just user satisfaction but user experience as a whole. "You like it when it's easy for you, when you don't have to put in too much effort." Likeable in the sense that it's a pleasant experience considering you have a goal to be accomplished. Easy to use – automatically likeable? Not for everyone. There is a distinction between efficient and likeable, as long as you present the user with both options. Likeable on its own appears to be interpreted more holistically, too broad. Likeability applies more to which one you would go back to and use again?
- "I found the chatbot's responses clear" matched to understandability, appropriate language style, graceful responses, perceived credibility.
- "I find the chatbot easy to use" matched to four factors (6, 12, 14 and 15) – users agree that all of these should be separate and the general one is pretty redundant. Ease of use seems to cover many of these factors and each person has a different idea of what easy to use entails.
- "using hyperlinks" – reference to service, and recognition and facilitation of user's goal and intent

Advice on building the questionnaire: allow users to preview all questions so people could rate questions relative to each other. Should it be done like this? This isn't what we want though. Give descriptions but we'd be leading the user so ideally the user should be able to interpret them accurately with ease.

15 item questionnaire – max 20 items.

Clustering into larger groups. Efficient for finding out where exactly the fault lies. Again might lead the user.

Questions should be more to the point.

1.7. Number of participants that assigned an item to more than one factor (in descending order)

Item	Factor	P11	P12	P13	P14	P21	P22	P23	P24	P31	P32	P33	P34	P41	P42	P43	P44	Multiple factor assignment
41	19	1	0	1	0	1		1	1	1	0	1	0	1	1	1	1	11
7	7	0	0	1	0	1	1	0	0	1	1	0	1	0	1	1	1	9
28	7	1	1	0	1	0		0	1	0	1	0	0	0	1	1	1	8
5	5	1	0	1	0	0	1	0	0	0	1	1	1	0	1	0	0	7
6	6	1	0	0	0	0	1	0	0	0	1	1	0	1	1	1	0	7
15	15	1	0	0	0	1	1	0	1	0	0	1	0	1	1	0		7
35	14	1	0	0	0	1		1	1	0	0	1	1	1	0	0	0	7
51	10	1	0	1	0	0		0	0	0	0	1	0	1	1	1	1	7
53	12	1	0	0	0	1			1	0	1	0	0	1	1	0	1	7
9	9	0	0	0	0	0	1	0	0	0	1	1	1	1	1	0	0	6
12	12	1	0	0	0	0	1	0	0	0	1	1	0	1	1	0	0	6
19	19	0	0	1	0	0		0	0	0	1	1	0	1	1	1	0	6
33	12	1	0	0	0	1		1	0	1	0	1	0	1	0	0	0	6
59	19	1	1	1	0	1			0	0	0	0	1	0	1	0	0	6
26	5	0	1	1	0	0		0	0	1	0	1	0	0	1	0	0	5
39	18	1	0	0	0	0		1	0	0	0	0	0	1	1	0	1	5
55	14	1	1	0	0	1			0	0	0	1	0	1	0	0	0	5
1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	4
18	18	0	0	0	0	0		0	0	0	0	1	1	0	1	0	1	4
22	1	1	0	0	0	0		0	0	0	0	1	0	0	1	0	1	4

43	1	1	0	0	0	1		1	0	0	0	0	0	0	0	0	1	4
46	5	0	0	0	0	0		0	1	1	0	1	0	0	1	0	0	4
58	18	1	0	0	0	0			0	1	0	0	0	1	0	1	0	4
4	4	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	3
10	10	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	3
14	14	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	3
16	16	0	0	0	0	0	1	0	0	0	0	1	0		1	0	0	3
47	6	1	0	0	0	1		0	0	0	0	0	0	0	1	0	0	3
48	7	0	0	0	0	0		0	0	0	0	0	0	1	1	0	1	3
50	9	0	0	0	0	1		0	0	0	0	1	0	0	1	0	0	3
25	4	0	0	0	0	0		0	0	1	0	0	0	0	1	0	0	2
30	9	1	0	0	0	0		0	0	0	0	0	0	0	1	0	0	2
37	16	0	0	0	0	0		0	0	0	0	0	0	0	1	1	0	2
56	15	1	0	0	0	0			0	0	0	0	0	0	1	0	0	2
27	6	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	1
31	10	0	0	0	0	0		0	0	0	0	0	0	0	1	0	0	1
45	4	1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	1
13	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	13	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0
36	15	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0
54	13	0	0	0	0	0			0	0	0	0	0	0	0	0	0	0

Appendix 2

2.1. Chatbots and tasks

Pre-tested chatbots

1. https://www.ato.gov.au/

Task 1: 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.

2. https://www.amtrak.com/home

Task 1: You have planned a trip to the USA. You are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet an old friend for a few hours and see the city. You want to use Amtrak's chatbot to find out how much it will cost to temporarily store your luggage at the station.

3. https://www.inbenta.com/en/

Task 1: You have an interview with Inbenta in a few days and you want to use Inbenta's chatbot to find out the address of Inbenta's Mexico office.

4. http://www.toshiba.co.uk/generic/yoko-home/

Task 1: You have Toshiba laptop of Satellite family and you are using Windows 7 operating system on your laptop. You want to partition your hard drive because it will make it easier to organize your video & audio libraries.

New chatbots

5. https://www.uscis.gov/emma

Task: You are a US citizen living abroad and want to vote in the upcoming federal elections. You want to use the USCIS chatbot to find out how.

6. https://www.facebook.com/messages/t/131840030178250 **(booking.com)**

Task: You are travelling to London from 5th July to 9th July with your family. You want to use booking.com's chatbot to find a hotel room for you, your significant other and your child in Central London that does not cost more than 500€ in total.

7. https://www.facebook.com/messages/t/1800FlowersAssistant **(1-800-FlowersAssistant)**

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

8. https://www.hsbc.co.uk/ **(HSBC UK)**

Task: You live in the Netherlands but are travelling to Turkey for 2 weeks. During your travel, you would like to be able to use your HSBC credit card overseas at payment terminals and ATMS. You want to use HSBC's chatbot to find out the relevant procedure.

9. https://www.absolut.com/en/ **(Absolut)**

Task: You want to buy a bottle of Absolut vodka to share with your friends for the evening. One of your friends cannot consume gluten. You want to use Absolut's chatbot to find out if Absolut Lime contains gluten or not.

10. **m.me/tommyhilfiger** **(Tommy Hilfiger)**

Task: You bought a perfume from a Tommy Hilfiger store in Paris for your friend. You have just gotten home (in the Netherlands) and found out that your friend already owns it. You want to use Tommy Hilfiger's chatbot to find out how to return it.

2.2 Session script

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

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

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

<<Give participant informed consent form>>

First, please fill in the demographic questionnaire.

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

Do you have any questions? Are you ready to start?

If so, you may begin with the first chatbot. Follow the instructions on the screen and if you have questions, you may ask me.

<<Start recording the screen>>

2.3. Qualtrics survey flow

- Standard: Condition (2 Questions)
- Standard: Demographics (7 Questions)
- Branch: New Branch
  - If
    - If Participant condition (for researcher only) B Is Selected
  - BlockRandomizer: 2 -
    - Standard: Amtrak (7 Questions)
    - Standard: Toshiba (7 Questions)
- Branch: New Branch
  - If
    - If Participant condition (for researcher only) A Is Selected
  - BlockRandomizer: 2 -
    - Standard: ATO (7 Questions)
    - Standard: Inbenta (7 Questions)
- BlockRandomizer: 3 - Evenly Present Elements
  - Standard: Flowers (7 Questions)
  - Standard: HSBC (7 Questions)
  - Standard: Absolut (7 Questions)
  - Standard: Booking.com (7 Questions)
  - Standard: USCIS (7 Questions)
  - Standard: Tommy Hilfiger (7 Questions)
- Standard: End (1 Question)

Page Break

2.4. Example of survey structure for a single chatbot

Start of Block: Condition

Q87 Participant ID

_____

Q13 Participant condition (for researcher only)

- ○ A (1)
- ○ B (2)

End of Block: Condition

Start of Block: Demographics

Gender Gender

▼ Male (1) ... Prefer not to say (3)

Age Age

_____

Nationality Nationality

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

Study Field of study

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

Familiarity

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

Prior_Usage

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

Display This Question:

If Prior_Usage = Definitely yes

Or Prior_Usage = Probably

Or Prior_Usage = Unsure

How_often

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

End of Block: Demographics

Start of Block: Amtrak

Amtrak

Chatbot: Amtrak

The chatbot can be found at: https://www.amtrak.com/home

Please access the chatbot now.

Page Break

Amtrak_Task

Please do the following task on this chatbot.

You have planned a trip to the USA. You are planning to travel by train from Boston to Washington D.C. You want to stop at New York to meet an old friend for a few hours and see the city. You want to use Amtrak's chatbot to find out how much it will cost to temporarily store your luggage at the station.

Page Break

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

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
It was clear how to start a conversation with the chatbot. (1)	○	○	○	○	○
It was easy for me to understand how to start the interaction with the chatbot. (2)	○	○	○	○	○
I find it easy to start a conversation with the chatbot. (3)	○	○	○	○	○
The chatbot was easy to access. (4)	○	○	○	○	○
The chatbot function was easily detectable. (5)	○	○	○	○	○
It was easy to find the chatbot. (6)	○	○	○	○	○
Communicating with the chatbot was clear. (7)	○	○	○	○	○
I was immediately made aware of what information the chatbot can give me. (8)	○	○	○	○	○
It is clear to me early on about what the chatbot can do. (9)	○	○	○	○	○
I had to rephrase my input multiple times for the chatbot to be able to help me. (10)	○	○	○	○	○

I had to pay special attention regarding my phrasing when communicating with the chatbot. (11)	○	○	○	○	○
It was easy to tell the chatbot what I would like it to do. (12)	○	○	○	○	○
The interaction with the chatbot felt like an ongoing conversation. (13)	○	○	○	○	○
The chatbot was able to keep track of context. (14)	○	○	○	○	○
The chatbot maintained relevant conversation. (15)	○	○	○	○	○
The chatbot guided me to the relevant service. (16)	○	○	○	○	○
The chatbot is using hyperlinks to guide me to my goal. (17)	○	○	○	○	○
The chatbot was able to make references to the website or service when appropriate. (18)	○	○	○	○	○
The interaction with the chatbot felt secure in terms of privacy. (19)	○	○	○	○	○

I believe the chatbot informs me of any possible privacy issues. (20)	○	○	○	○	○
I believe that this chatbot maintains my privacy. (21)	○	○	○	○	○
I felt that my intentions were understood by the chatbot. (22)	○	○	○	○	○
The chatbot was able to guide me to my goal. (23)	○	○	○	○	○
I find that the chatbot understands what I want and helps me achieve my goal. (24)	○	○	○	○	○
The chatbot gave relevant information during the whole conversation (25)	○	○	○	○	○
The chatbot is good at providing me with a helpful response at any point of the process. (26)	○	○	○	○	○
The chatbot provided relevant information as and when I needed it. (27)	○	○	○	○	○
The amount of received information was neither too much nor too less (28)	○	○	○	○	○

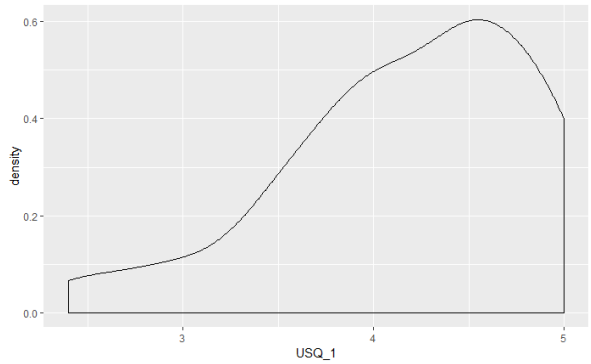
The chatbot gives me the appropriate amount of information (29)	○	○	○	○	○
The chatbot only gives me the information I need (30)	○	○	○	○	○
The chatbot could handle situations in which the line of conversation was not clear (31)	○	○	○	○	○
The chatbot explained gracefully when it could not help me (32)	○	○	○	○	○
When the chatbot encountered a problem, it responded appropriately (33)	○	○	○	○	○
I found the chatbot's responses clear. (34)	○	○	○	○	○
The chatbot only states understandable answers. (35)	○	○	○	○	○
The chatbot's responses were easy to understand. (36)	○	○	○	○	○
I feel like the chatbot's responses were accurate. (37)	○	○	○	○	○
I believe that the chatbot only states reliable information. (38)	○	○	○	○	○

It appeared that the chatbot provided accurate and reliable information. (39)	○	○	○	○	○
The time of the response was reasonable. (40)	○	○	○	○	○
My waiting time for a response from the chatbot was short. (41)	○	○	○	○	○
The chatbot is quick to respond. (42)	○	○	○	○	○

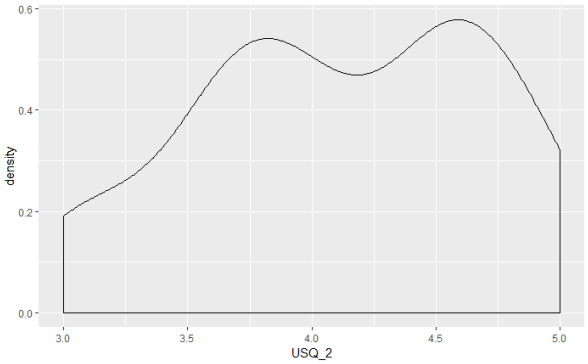
2.5. Item evaluation statistics

Item	Factor	mean	var	min	max	range	skew	kurtosis	se	r.drop
1	1	4.17	0.423	2.4	5	2.6	-0.77	0.03	0.09	0.634
10	1	3.06	0.436	1.2	4.2	3	-0.53	-0.11	0.09	*0.264*
11	1	3.25	0.578	1.6	4.75	3.15	-0.01	-0.72	0.1	*0.274*
2	1	*4.11*	0.348	3	5	2	-0.24	-1.03	0.08	0.649
3	1	*4.2*	0.240	3.25	5	1.75	-0.03	-1.13	0.07	0.637
4	1	*4.15*	0.250	3.2	5	1.8	-0.08	-1.1	0.07	0.568
5	1	3.86	0.325	2.67	5	2.33	-0.14	-0.81	0.08	0.773
6	1	3.97	0.270	2.67	5	2.33	-0.23	-0.52	0.07	0.692
12	2	3.53	0.436	2.2	4.8	2.6	0.04	-0.86	0.09	0.579
14	2	3.18	0.410	2	4.25	2.25	-0.27	-1.07	0.08	0.555
15	2	3.33	0.303	1.5	4.25	2.75	-0.8	0.88	0.07	0.594
16	2	3.65	0.462	1.75	5	3.25	-0.67	0.31	0.09	0.791
17	2	3.84	0.504	1.75	5	3.25	-0.37	-0.44	0.09	0.585
18	2	3.86	0.533	2	5	3	-0.5	-0.26	0.1	0.668
22	2	3.29	0.397	2	4.4	2.4	-0.29	-0.86	0.08	0.772
23	2	3.59	0.462	1.8	4.8	3	-0.36	-0.17	0.09	0.750
24	2	3.38	0.410	1.8	4.8	3	-0.23	-0.23	0.08	0.858
25	2	3.27	0.397	1.8	4.5	2.7	-0.14	-0.8	0.08	0.855
26	2	3.33	0.360	1.8	4.6	2.8	-0.38	-0.12	0.08	0.859
27	2	3.38	0.325	1.8	4.6	2.8	-0.2	-0.16	0.08	0.844
28	2	3.28	0.476	1.8	4.5	2.7	-0.23	-0.71	0.09	0.593
29	2	3.36	0.360	2	4.5	2.5	-0.3	-0.61	0.08	0.719
30	2	3.04	0.423	1.75	4.33	2.58	0.15	-0.6	0.09	0.763
31	2	2.84	0.410	1.6	4.8	3.2	0.19	0.17	0.08	0.579
32	2	3.11	0.397	1.6	4.67	3.07	-0.03	0.01	0.08	0.340
33	2	3.13	0.314	2	4.6	2.6	0.11	-0.43	0.07	0.620
34	2	3.58	0.436	1.8	4.6	2.8	-0.61	-0.24	0.09	0.706
35	2	3.72	0.348	2	5	3	-0.51	0.38	0.08	0.482
36	2	3.86	0.292	2.8	5	2.2	-0.08	-0.84	0.07	0.600
37	2	3.46	0.436	1.6	4.8	3.2	-0.44	-0.02	0.09	0.707
38	2	3.62	0.348	2.2	5	2.8	-0.05	-0.49	0.08	0.654
39	2	3.73	0.360	2.4	5	2.6	0.01	-0.65	0.08	0.703
7	2	3.71	0.314	2.4	5	2.6	0	-0.43	0.07	0.602
8	2	3.36	0.476	2	4.6	2.6	-0.24	-0.93	0.09	0.512
9	2	3.53	0.548	1.75	5	3.25	-0.42	-0.78	0.1	0.408
19	3	3.12	0.608	1	4.6	3.6	-0.4	-0.4	0.1	0.816
20	3	2.49	0.792	1	4.6	3.6	0.31	-0.47	0.12	0.735
21	3	2.95	0.563	1.25	4.6	3.35	-0.12	-0.5	0.1	0.826
40	4	*4.29*	0.303	2.6	5	2.4	-0.59	-0.15	0.07	0.740
41	4	*4.3*	0.281	3.2	5	1.8	-0.44	-0.82	0.07	0.898
42	4	*4.26*	0.325	2.33	5	2.67	-0.67	0.49	0.08	0.849

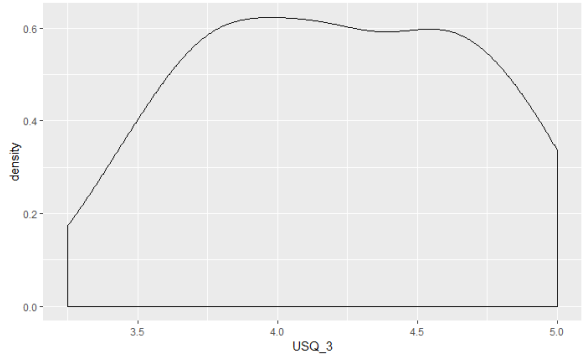
2.6. Item histograms



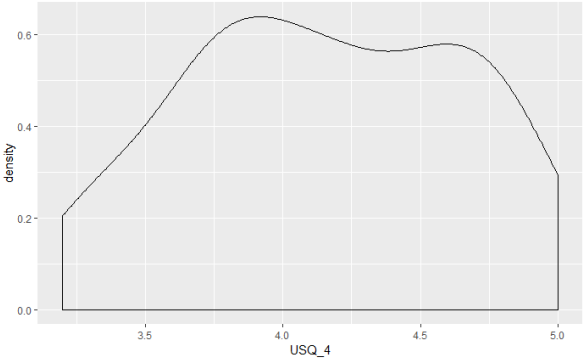
Y1



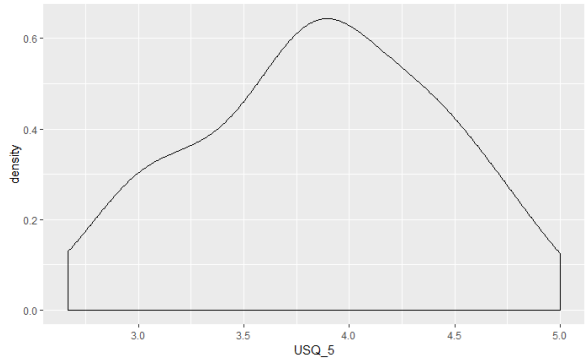
Y2



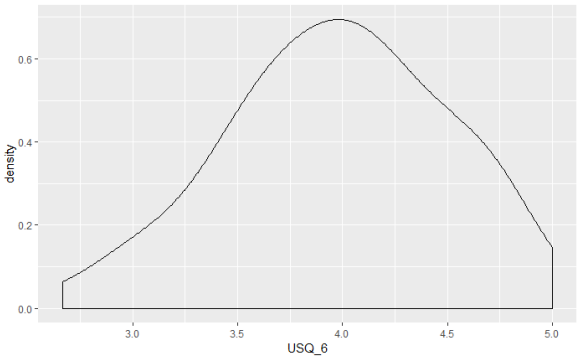
Y3



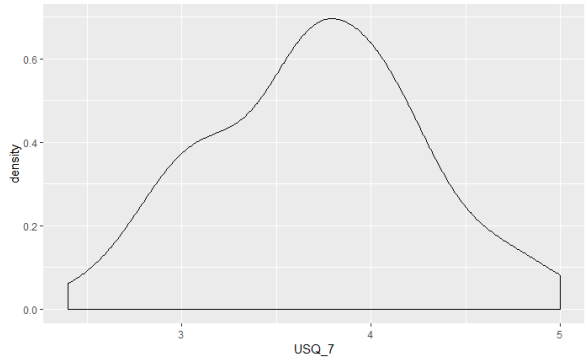
Y4



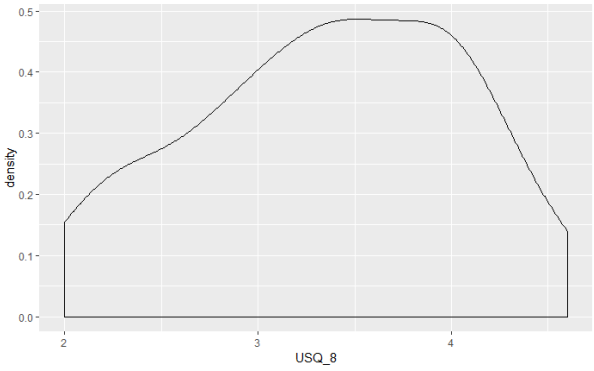
Y5



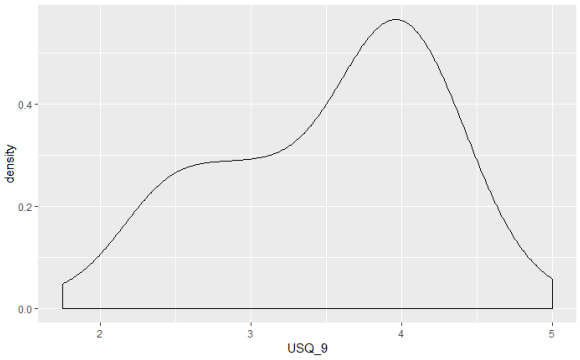
Y6



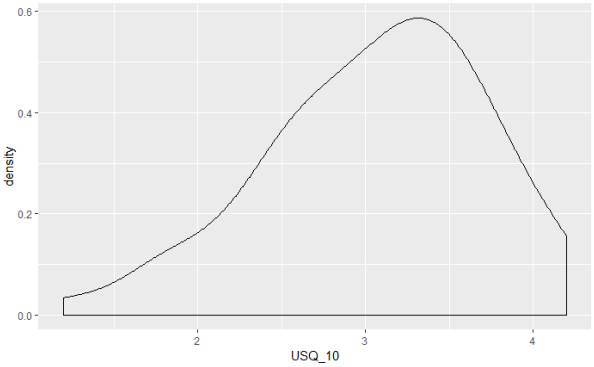
Y7



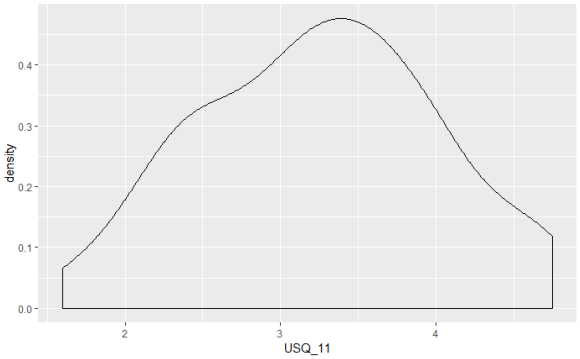
Y8



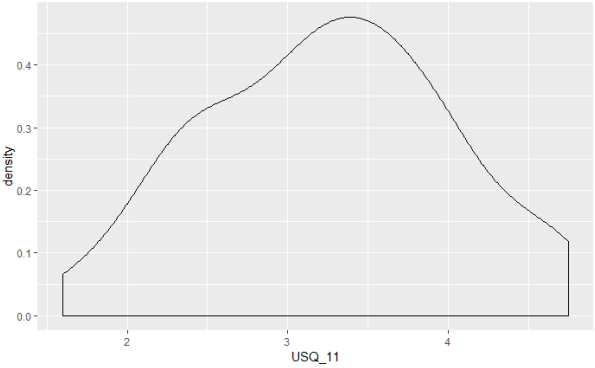
Y9



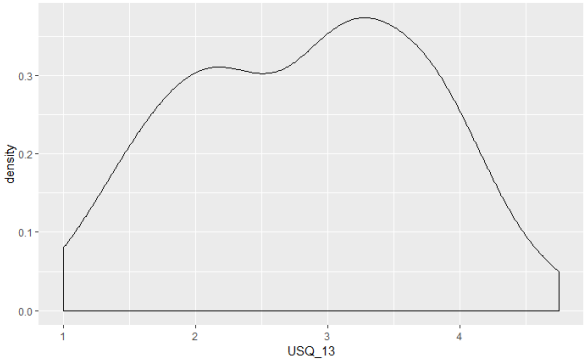
Y10



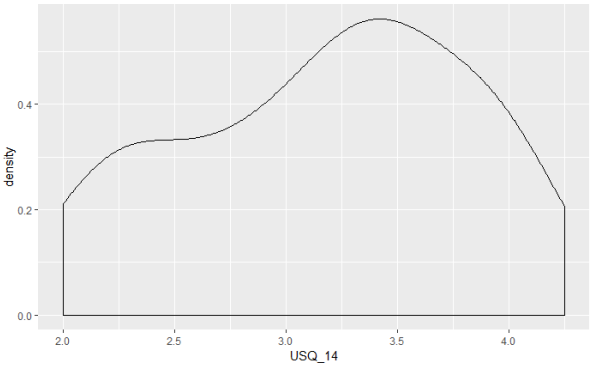
Y11



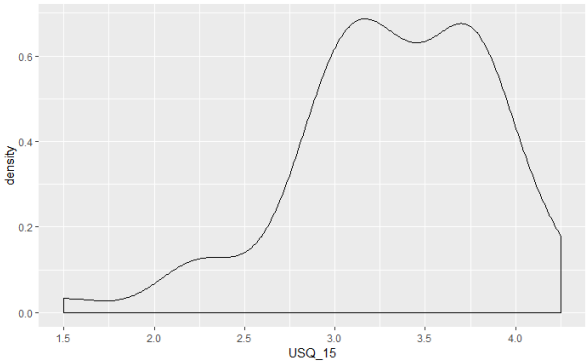
Y12



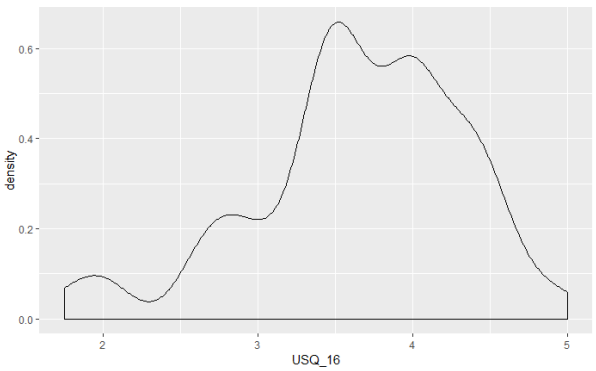
Y13



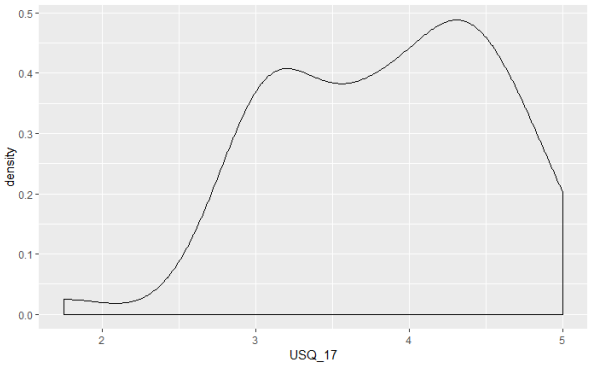
Y14



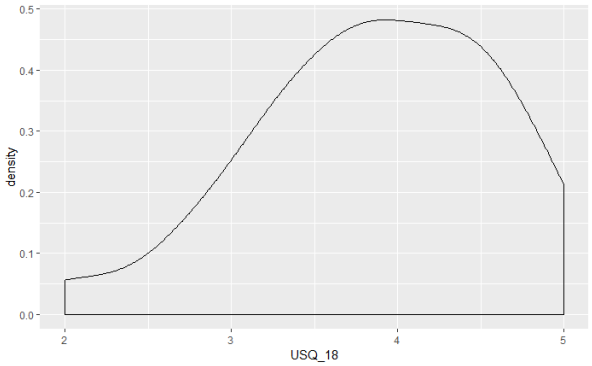
Y15



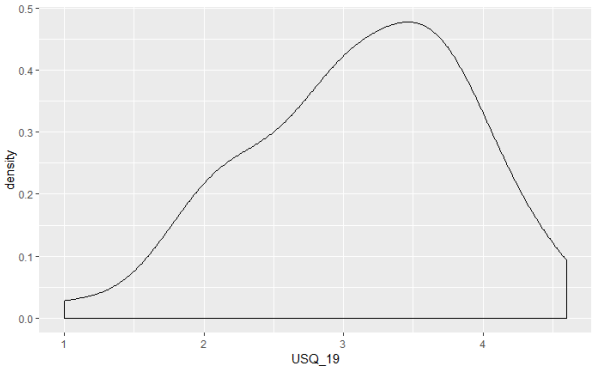
Y16



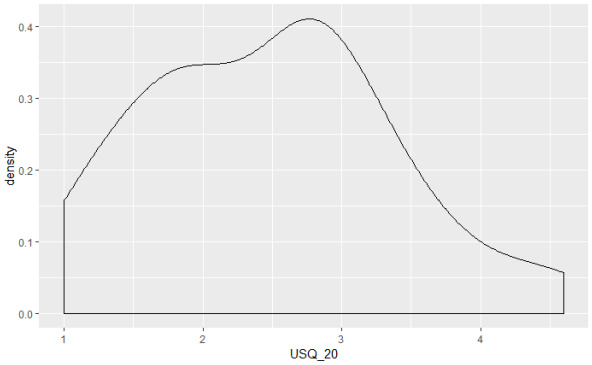
Y17



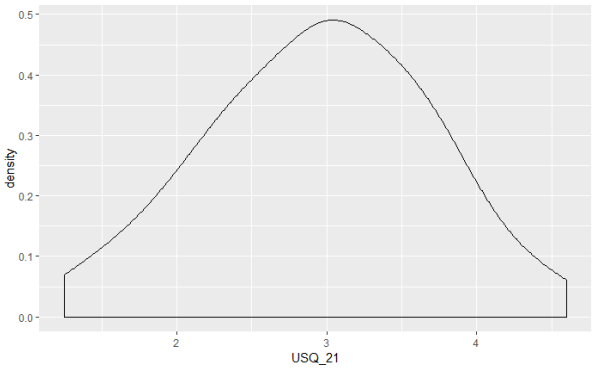
Y18



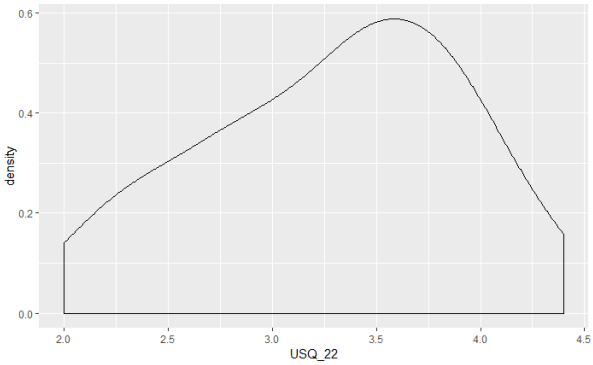
Y19



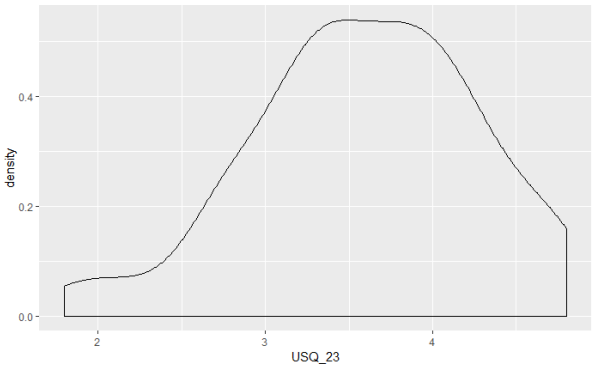
Y20



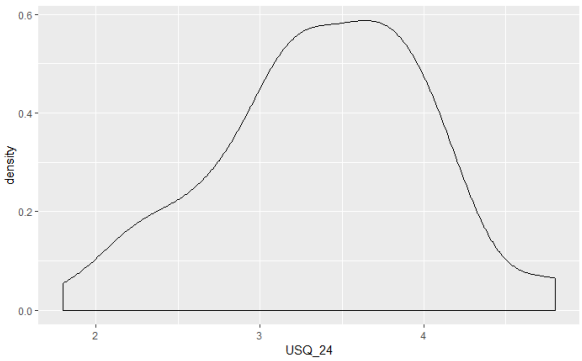
Y21



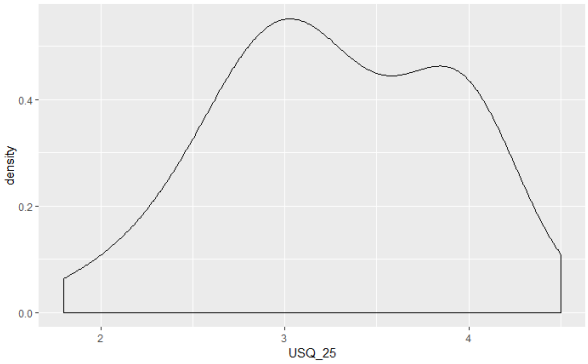
Y22



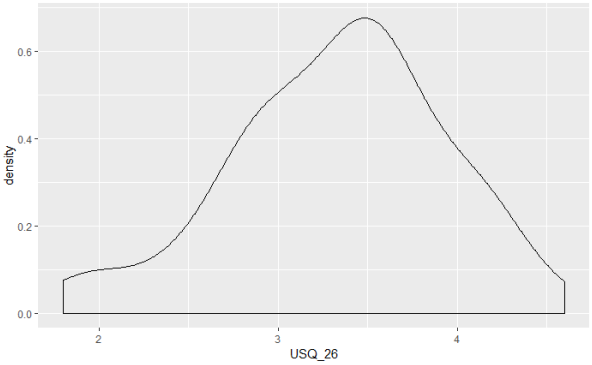
Y23



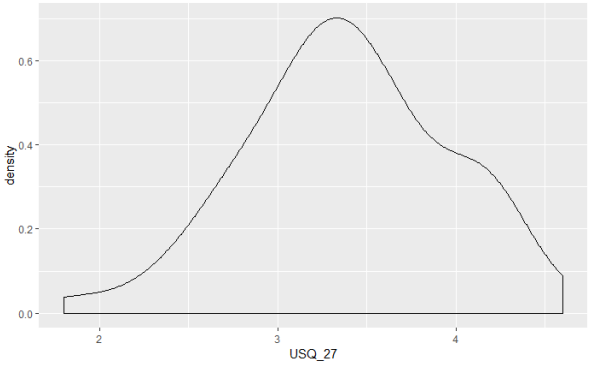
Y24



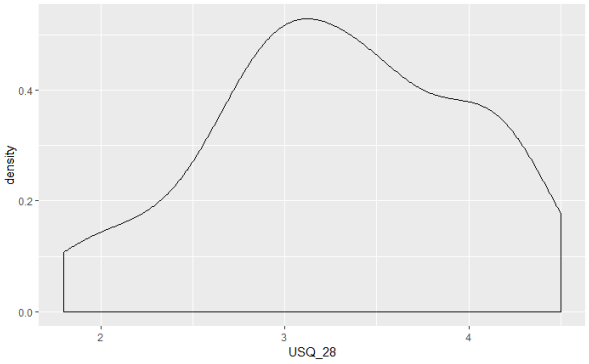
Y25



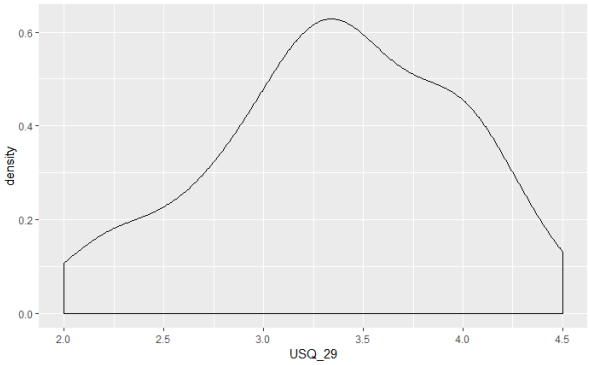
Y26



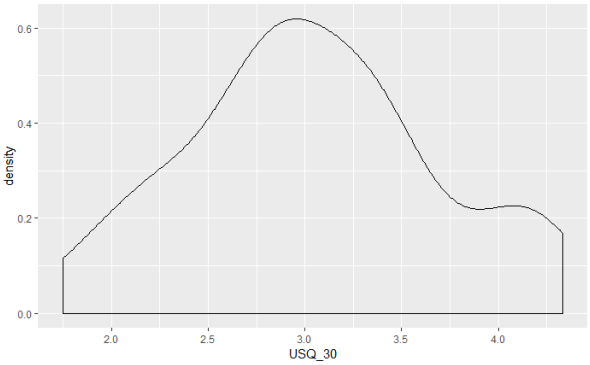
Y27



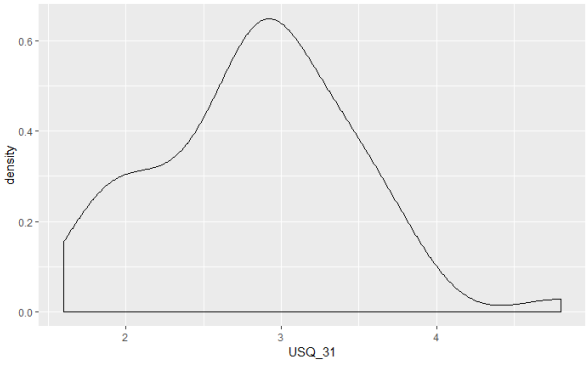
Y28



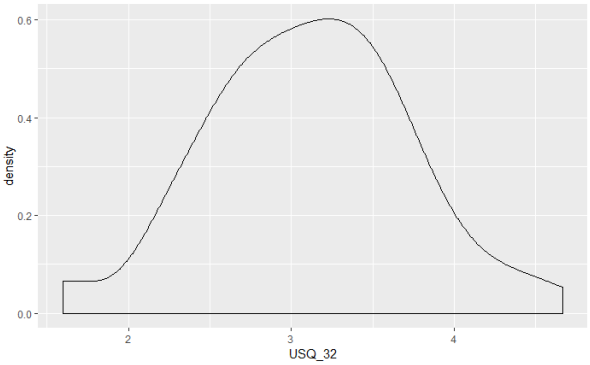
Y29



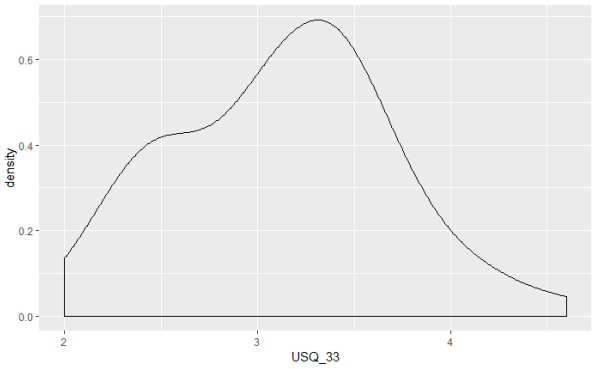
Y30



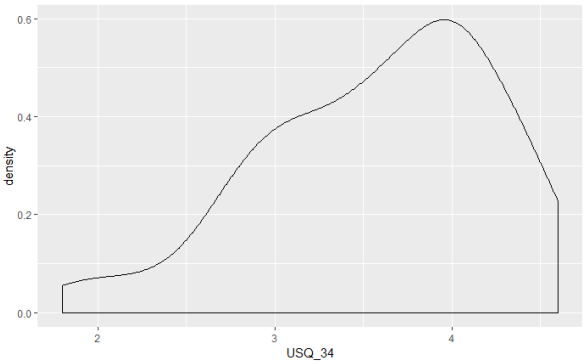
Y31



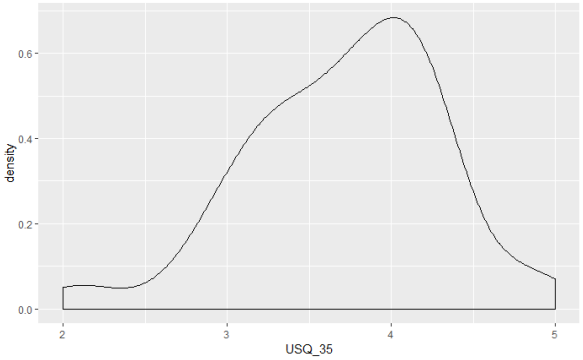
Y32



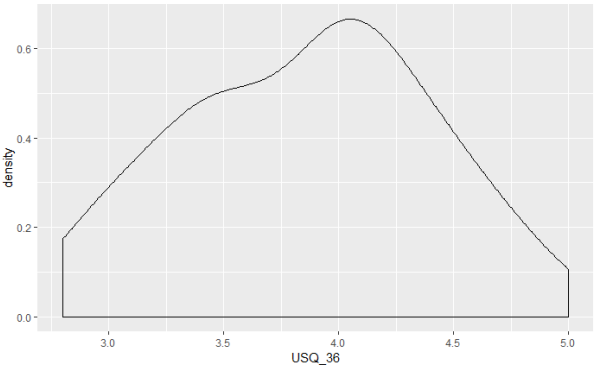
Y33



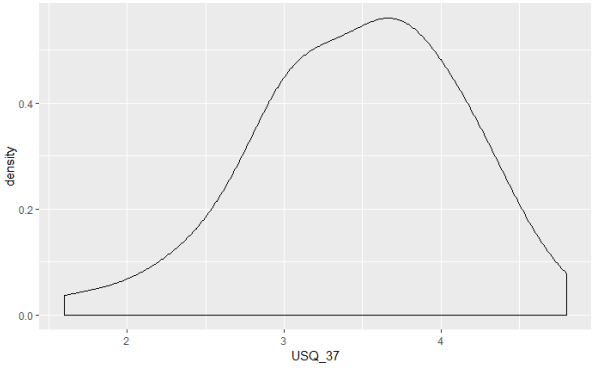
Y34



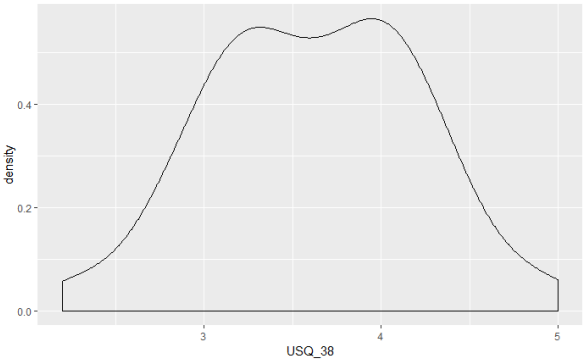
Y35



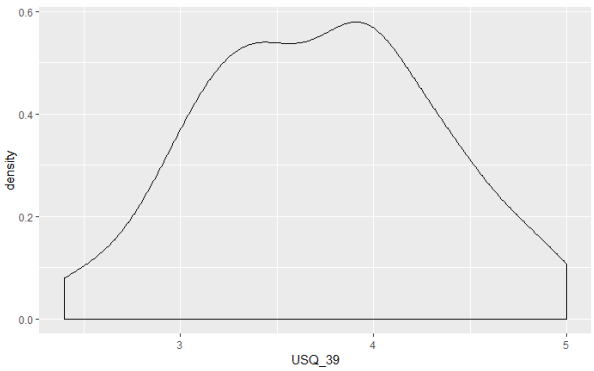
Y36



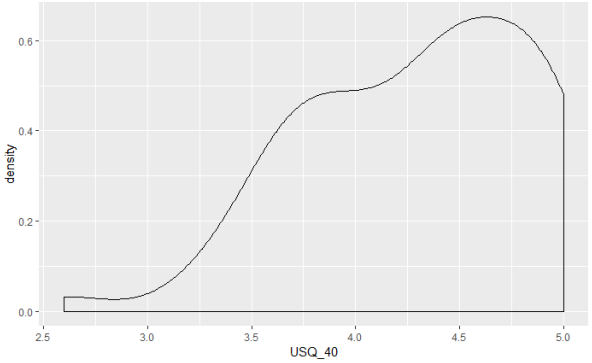
Y37



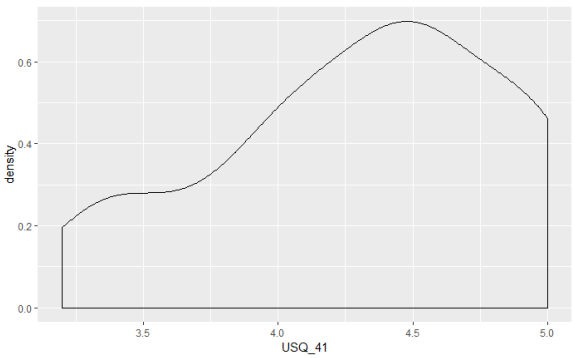
Y38



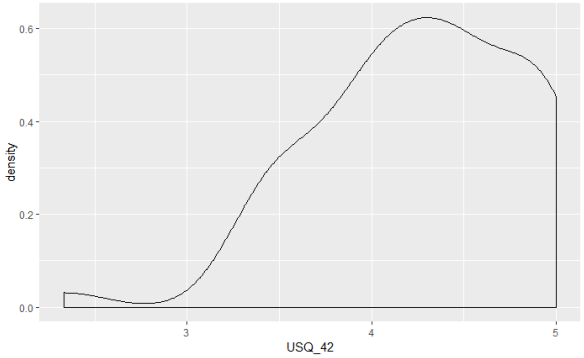
Y39



Y40



Y41



Y42

2.7. R code used to run analyses in Study 2

Libraries

```
library(tidyverse)

library(dplyr)
library(knitr)
library(DT)
library(xtable)

library(psych)

library(GPArotation)

library(blavaan)

library(rstan)

library(BayesFM)
```

Importing dataset

Rows containing missing data were removed. Participant column also removed. All variables mutated to 'double' for factor analysis.

```
chatbots <-
  read_csv("Thesis Analysis/Chatbots2.csv")

## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Part = col_integer()
## )

## See spec(...) for full column specifications.

fa <-
  chatbots[-c(49, 50, 52), -c(1)]%>%
  mutate_all(as.double)
```

Factor Analysis

Changing column names of dataset for referencing purposes.

```
colnames(fa) <-
  c("x1", "x10", "x11", "x12", "x13", "x14", "x15", "x16", "x17", "x18", "
x19", "x2", "x20", "x21", "x22", "x23", "x24", "x25", "x26", "x27", "x28",
"x29", "x3", "x30", "x31", "x32", "x33", "x34", "x35", "x36", "x37", "x38"
, "x39", "x4", "x40", "x41", "x42", "x5", "x6", "x7", "x8", "x9")
```

Specifying hypothesized models for testing

- Model 1: 8 factors
- Model 2: 5 factors

```
model1 <- '
  Content relevance =~ x25 + x26 + x27 + x22 + x23 + x24 + x13 + x14 + x15
+ x37 + x38 + x39
  Response clarity =~ x28 + x29 + x30 + x34 + x35 + x36
  Speed =~ x40 + x41 + x42
  Graceful breakdown =~ x31 + x32 + x33
  Reference to service =~ x16 + x17 + x18
  Initiating conversation =~ x4 + x5 + x6 + x1 + x2 + x3 + x7 + x8 + x9
  Communication effort =~ x10 + x11 + x12
  Perceived privacy =~ x19 + x20 + x21'

model2 <- '
  Communication quality =~ x4 + x5 + x6 + x1 + x2 + x3 + x7 + x8 + x9 + x1
0 + x11 + x12
  Speed =~ x40 + x41 + x42
  Perceived privacy =~ x19 + x20 + x21
  Graceful breakdown =~ x31 + x32 + x33
  Interaction quality =~ x25 + x26 + x27 + x22 + x23 + x24 + x13 + x14 + x
15 + x37 + x38 + x39 + x28 + x29 + x30 + x34 + x35 + x36 + x16 + x17 + x18
'
```

Confirmatory Factor Analysis

Testing model 1

```
  fit1 <-
  bcfa(model1,
       data=fa,
       target="stan",
       control = list(adapt_delta=0.99, stepsize = 0.001, max_treedepth =
15),
       cp = "srs",
       n.chains = 3,
       test="none")

summary(fit1)
```

Testing model 2

```
fit2 <-
  bcfa(model2,
       data=fa,
```

```
       target="stan",
       control = list(adapt_delta=0.99, stepsize = 0.001, max_treedepth =
15),
       cp = "srs",
       n.chains = 3,
       test="none")

summary(fit2)
```

Parallel analysis

Results suggest number of factors is 3. Inspection of scree plots suggests number of factors lies between 3 and 7.

```
nofactors = fa.parallel(fa, fm="ml", fa="fa")
```

Exploratory factor analysis

7 factors

```
fa7 <-
  befa(fa,
       burnin=5000,
       iter=50000,
       Kmax=7)

fa7 <- post.column.switch(fa7)
fa7 <- post.sign.switch(fa7)

fa7_hppm <-
  summary(fa7, what='hppm', min.prob=0.1)

fa7_hppm
```

4 factors

```
fa4 <-
  befa(fa,
       burnin=5000,
       iter=50000,
       Kmax=4)

fa4 <- post.column.switch(fa4)
fa4 <- post.sign.switch(fa4)

fa4_hppm <-
  summary(fa4, what='hppm')

fa4_hppm
```

3 factors

```
fa3 <-
  befa(fa,
       burnin=5000,
       iter=50000,
```

```
       Kmax=3)

fa3 <- post.column.switch(fa3)
fa3 <- post.sign.switch(fa3)

fa3_hppm <-
  summary(fa3, what='hppm')

fa3_hppm
```

Item evaluation statistics

Descriptive statistics

```
summary <-
  describe(fa)
```

Item histograms

```
plotHistFunc <-
  function(fa, na.rm = TRUE, ...) {
  nm <- colnames(fa)
  for (i in seq_along(nm)) {
  plots <- ggplot(fa, aes_string(x = nm[i])) + geom_density() +
    xlim(1,5)
  print(plots)
}
  }

plotHistFunc(fa)
```

Reliability analysis

```
CQ <- select(fa, 1, 12, 23, 34, 38, 39, 2, 3)
IQ <- select(fa, 40, 41, 42, 4, 6, 7, 8, 9, 10, 15, 16, 17, 18, 19, 20, 21
, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33)
PP <- select(fa, 11, 13, 14)
RS <- select(fa, 35, 36, 37)

psych::alpha(CQ, keys=c(1,1,1,1,-1,-1))
psych::alpha(IQ)
psych::alpha(PP)
psych::alpha(RS)

CQnew <- select(fa, 1, 12, 34, 38, 2, 3)
IQnew <- select(fa, 40, 7, 10, 17, 18, 24, 27, 28, 31)
```