




# The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents

Simone Borsci<sup>1,2</sup> · Alessio Malizia<sup>3,4</sup> · Martin Schmettow<sup>1</sup> · Frank van der Velde<sup>1</sup> · Gunay Tariverdiyeva<sup>5</sup> · Divyaa Balaji<sup>6</sup> · Alan Chamberlain<sup>7</sup> 

Received: 22 November 2020 / Accepted: 29 May 2021 / Published online: 21 July 2021  
© The Author(s) 2021

## Abstract

Standardised tools to assess a user's satisfaction with the experience of using chatbots and conversational agents are currently unavailable. This work describes four studies, including a systematic literature review, with an overall sample of 141 participants in the survey (experts and novices), focus group sessions and testing of chatbots to (i) define attributes to assess the quality of interaction with chatbots and (ii) the designing and piloting a new scale to measure satisfaction after the experience with chatbots. Two instruments were developed: (i) A diagnostic tool in the form of a checklist (BOT-Check). This tool is a development of previous works which can be used reliably to check the quality of a chatbots experience in line with commonplace principles. (ii) A 15-item questionnaire (BOT Usability Scale, BUS-15) with estimated reliability between .76 and .87 distributed in five factors. BUS-15 strongly correlates with UMUX-LITE by enabling designers to consider a broader range of aspects usually not considered in satisfaction tools for non-conversational agents, e.g. conversational efficiency and accessibility, quality of the chatbot's functionality and so on. Despite the convincing psychometric properties, BUS-15 requires further testing and validation. Designers can use it as a tool to assess products, thus building independent databases for future evaluation of its reliability, validity and sensitivity.

**Keywords** Artificial intelligence · AI · Chatbots · Conversational agents · Interaction satisfaction · Usability · User experience · Human-Computer interaction (HCI) · Evaluation · Autonomy · Design · Satisfaction · Trust

## 1 Introduction

Chatbots can be defined as intelligent conversational applications that can simulate natural language conversation by en-

gaging in text or voice (or both) input and output exchange with humans [56]. These tools may be designed to perform in different contexts (web platforms, social networks, home devices etc.) and to serve a wide range of goals in different domains from entertainment to health assistance and customer service support [4, 28]. As suggested by Radziwill and Benton [69]: 'chatbots are one class of intelligent, conversational software agents activated by natural language input'. Conversational agents are generally categorised as highly driven by artificial intelligence while chatbots could be more or less sophisticated in their ability to drive the natural conversation with end-users or to help customers in achieving their goals. Nevertheless, in literature chatbots and conversational agents are often used as synonyms [42, 77]. When attached to a company service, chatbots aim to support the decision-making and information retrieval of end-users [62] and are generally used as customer relationship management (CRM) tools. These CRM chatbots may be used to reduce operational costs associated with customer service and to enhance the brand image by providing 24/7 rapid and effective exchanges with costumers to facilitate the access to

✉ Alan Chamberlain  
alan.chamberlain@nottingham.ac.uk

<sup>1</sup> Department of Learning, Data analysis, and Technology, Cognition, Data and Education (CODE) group, Faculty of Behavioural Management and Social sciences, University of Twente, Enschede, Netherlands

<sup>2</sup> NIHR London In-Vitro Diagnostics Cooperative, Imperial College of London, London, UK

<sup>3</sup> Computer Science Department, University of Pisa, Pisa, Italy

<sup>4</sup> Molde University College, Molde, Norway

<sup>5</sup> Backbase, Amsterdam, Netherlands

<sup>6</sup> Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, Netherlands

<sup>7</sup> School of Computer Science, University of Nottingham, Nottingham, United Kingdom

information [28]. These service tools, being usually proprietary or customised systems of a company, may substantially vary in terms of appearance, behaviour and capabilities and provide a different experience to end-users [17].

Forecasting data suggest that around the 85% of customer interactions will be handled without a human agent by 2020 with an expected market value of conversational agents of \$ 6 Billion by 2023 [7].

Despite the potential market for conversational agents, Valério et al. [78] suggest there is still too little known about how to assess the end-user perception of quality when interacting with chatbots. Evaluation frameworks such as the PARADigm for Dialogue System Evaluation (PARADISE, [82]) suggests the end-users satisfaction with chatbots should be considered as a weighted product of success in achieving the tasks (maximise task success) at an acceptable cost (efficiency and quality of Chatbot's performance). In line with this paradigm, Radziwill and Benton [69] recently conducted a literature review and compiled a list of thirty-eight quality attributes which can be used to design conversational agents. These authors proposed a list of qualities attributed to Chatbots; these are intended to be used as guidelines (or checklists) for designers. In this work, we convert a design-oriented 'attributes list' into an inventory to measure satisfaction with chatbots. Satisfaction is a tricky measure of the end-user reaction and reasoning about systems which relates to efficiency and effectiveness and accurate and reliable modalities of assessment [3, 22, 29, 44, 54]. Nevertheless, as recognised by Thorne [76], researchers in the field of conversational agents tend to translate methods from Human-Computer Interaction (HCI) simply. Often satisfaction with chatbots is measured using tools developed to assess web or digital interfaces [71]. Reliable and short scales such as the System Usability Scale (SUS, [6]) and its shorter proxies the Usability Metric for User Experience (UMUX, [25]) and UMUX LITE [53] can undoubtedly guarantee comparable measures of satisfaction; however, these tools were not developed to consider the conversational aspects which relate to a user's interaction with conversational agents. Tools to assess speech user interface, voice respondents and voice controlled interface are available –e.g. Speech User Interface Service Quality scale [52, 67]; Mean Opinion Scale [50]; Subjective Assessment of Speech System Interfaces [38]. Such tools, however, focus on technologies that are significantly less interactive than artificial intelligence (or advanced algorithms) based chatbots for CRM. The ability to communicate and maintain an efficient and effective conversational exchange is not a secondary, but actually, a characterising element of chatbots that should be considered in the assessment of satisfaction with these tools [18, 19]. By partially recognising this issue, some researchers used qualitative instruments that directly inquire about the overall impression/experience of the end-

users after a given interaction with chatbots; these assessments also take the conversational aspects of the experience into account [61, 68, 72]. The use of qualitative methods provides an insight into what constitutes a quality interactional experience of chatbot system, but such methods have not yet been translated into reliable and comparable instruments for assessment. As recently noted by Federici et al. [23], there is a growing need in the domain of chatbots and conversational agents to translate qualitative results into a validated scales to measure, diagnose and compare the quality of an chatbot-based interaction.

Attempts were made, in the marketing domain to systematise customer satisfaction toward a brand or a service that utilise conversational agents. For instance, in a recent marketing-oriented study [13], consumers of luxury brands with previous experience with chatbots assessed different agents by simply viewing screenshots of these systems to identify the benefit of using chatbots for marketing purposes. Concurrently, a recent work investigated the sources of satisfaction and dissatisfaction during the interaction with chatbots from the marketing perspective [86]. Moreover, it was recently proposed to use sentiment analysis as a way to automatically infer the sentiment toward a brand or a company after the exchange with a chatbot [24].

However, there is a difference between the satisfaction intended in the marketing domain as 'the customer's emotive post-consumption evaluation of the service performance' [80], which is inherently connected to the concept of loyalty [8], and the satisfaction of interaction defined in the ISO 9241-11 [43] as the 'extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations'. In the first case, the conversational agents are assessed to understand how to optimise the reaction toward a brand or a company service; in the latter chatbots are the object of observation and are evaluated to understand how to ameliorate the chatbots' performances to make the users satisfied of the interaction with the chatbot as a tool.

While we believe that the marketing and the interaction perspectives on satisfaction are complementary, the present work focuses on the latter by aiming at providing a toolkit to help designers of chatbots to consider during the design the needs of the end-users and to assess during the formative phase of development the satisfaction of the user with the interaction with a chatbot without considering the marketing implications that can and should be integrated at later stages of product development.

To the best of our knowledge, no previous studies have attempted to identify and test a model of users' satisfaction in the context of interaction with conversational agents by aiming at developing a reliable tool to guide the designers during the development.

To achieve the goal of developing tools to support designers in the evaluation of chatbots interactive quality, we performed four studies in sequence:

- i. The first study re-examines the attributes identified by Radziwill and Benton [69], based on a systematic review to identify attributes that end-users may indirectly or directly use to assess the quality of interaction after interacting with an information retrieval chatbots.
- ii. The second study was aimed at reaching consensus on this list of attributes. An online survey with chatbot designers and end-users was developed to accomplish this.
- iii. The third study aimed to expand the list attributes and to develop a list of ‘items’ for the questionnaire. Focus groups sessions were used to develop an initial version of the scale called the: Bot Usability Scale (BUS).
- iv. The goal of the fourth was to pilot the initial version of the BUS scale to explore its psychometric properties to create a final version of the scale for future analysis.

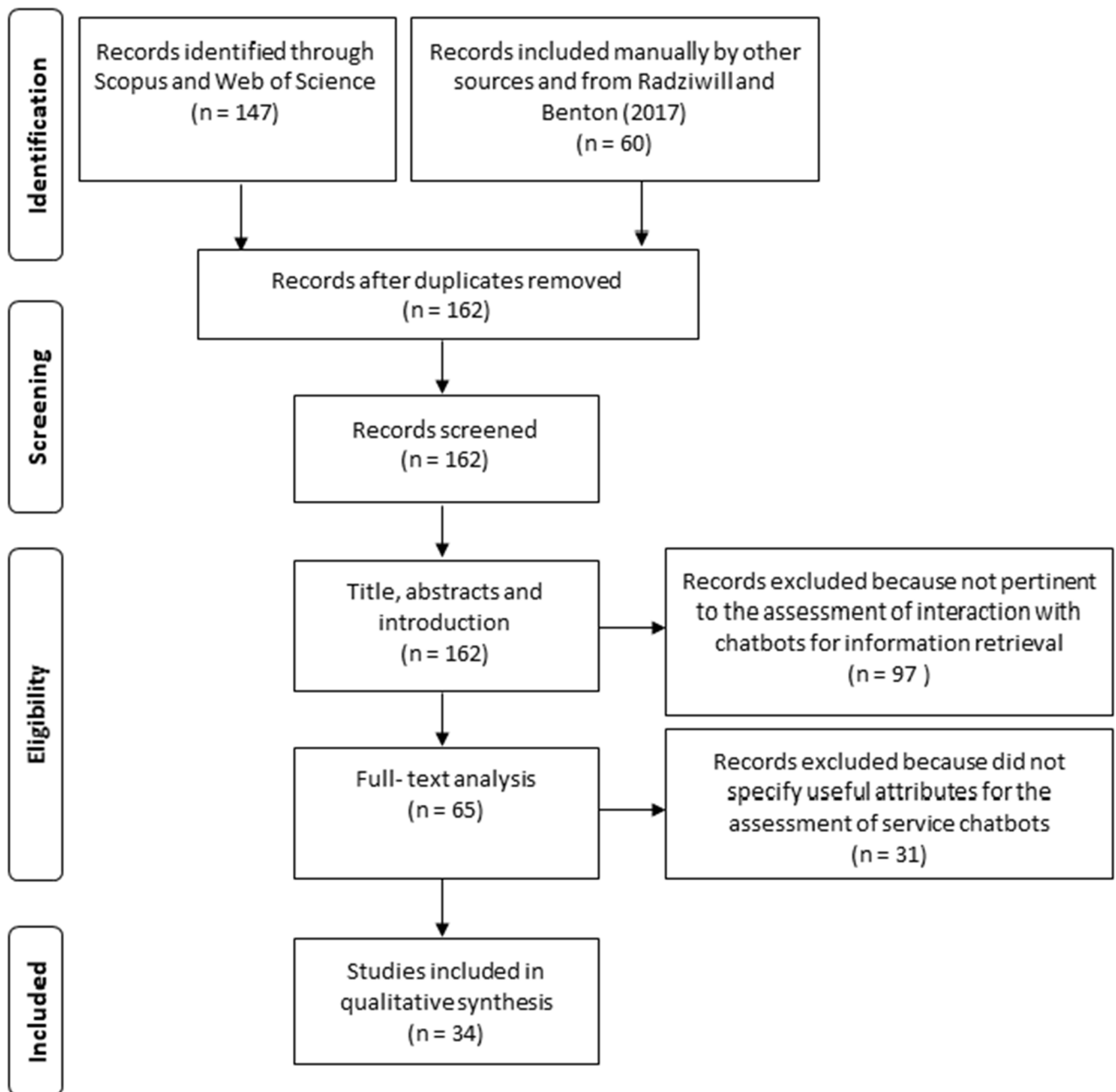


Fig. 1 PRISMA process of literature selection on the quality attributes of chatbots

## 2 Study 1—Attributes Collection

### 2.1 Methods

In line with Scale Development Theory [21, 75], we adopted a deductive approach to defining an initial construct in order to

assess the end-users satisfaction. Researchers screened and reviewed 26 references and 38 quality attributes proposed by Radziwill and Benton [69], which were initially developed as guidelines for designers. The goal of this screening was to identify attributes that could be used by end-users. During the re-examination of the literature and ‘quality-attributes’, attributes

**Table 1** The revised list of 27 attributes that can play a role in the end-users’ assessment of satisfaction after CRM chatbot–user interaction

Attribute code, name and descriptor	Reference List	Attributes compared with Radziwill and Benton [69]: New (N) Adapted (A) Retained (R)
A.1 Response Time. The chatbot is perceived as able to respond in a timely manner to requests	[1, 20, 40, 65]	N
A.2 Multi-thread conversation. The chatbot is perceived as able to recognise and process simultaneously multiple and parallel topics during the conversation	[49, 69, 74, 85]	A—Original description ‘Effective function allocation, provides appropriate escalation channels to humans’
A.3 Maxim of quantity. The chatbot responds in an informative way without adding too much information	[31, 33, 64]	N
A.4 Maxim of quality. The chatbot seems able to convey correct statements and information (perceived credibility)		N
A.5 Maxim of manners. The chatbot makes its purpose clear without ambiguity (understandability)		N
A.6 Maxim of relation. The chatbot provides a relevant and appropriate contribution to people’s needs at each stage		A - This attribute merged two attributes ‘Execute requested tasks’ and ‘Able to respond to specific questions’
A.7 Appropriate language style. Chatbot uses appropriate and accurate language style for the context	[46, 60]	A—This attribute merged two attributes ‘Appropriate degrees of formality’ and ‘Linguistic accuracy of outputs’
A.8 Reference to the service. Chatbot seems designed to use the environment (information, options, buttons on-screen, etc.) to guide the user towards its goal	[33]	N
A.9 Visual Look. The designed appearance of a chatbot’s dialogue box, avatar, font, etc.	[1, 47]	N
A.10 Voice Tone. The chatbot has an appropriate expressiveness (inflection, emotional information through tone) and accuracy of the text-to-speech function	[47, 63]	A—Original description ‘Provide emotional information through tone, inflection and expressivity’
A.11 Integration with the website or platform (visibility). The chatbot is located on the screen, it is visible and perceived as well integrated	[30, 47]	N
A.12 Graceful responses in unexpected situations. Chatbot seems able to handle gracefully unexpected events such as communication mismatch, or a broken line of conversation, etc.	[14, 34, 41, 45]	A—This attribute merged two attributes ‘Graceful degradation’ and ‘Robustness to manipulation’
A.13 Recognition and facilitation of users’ goal and intent. Chatbot seems able to recognise the user’s intent and guide the user to its goals	[15, 79, 84]	A—This attribute merged two attributes: chatbot ‘Can detect meaning or intent’ and ‘Interprets commands accurately’
A.14 Variation of responses. Chatbot seems able to respond in different and appropriate ways to similar or repeated requests	[33, 65]	N
A.15 Perceived ease of use. The interaction with the chatbot is perceived as free from errors	[69]	A—Original description ‘General ease of use’
A.16 Engage in on-the-fly problem-solving. Chatbot seems able to solve problems instantly on the spot	[4]	R
A.17 Ability to maintain a themed discussion. Chatbot maintains a conversational theme once introduced and keep track of the context to understand the user’s utterances	[33, 46, 47]	N
A.18 Breadth of knowledge. Chatbot seems able to exhibit knowledge that is out of its immediate domain during a conversation	[14, 81]	A—Original description ‘Contains breadth of knowledge, is flexible in interpreting’
A.19 Initiative. The chatbot is able to initiate conversation (or to offer cues) for further discussion by offering suggestions, etc.	[1, 33, 46–48, 83]	N
A.20 Personality. Chatbot conveys a personality by providing greetings, self-introductory, empathy, information, etc.	[1, 46–48]	N

**Table 1** (continued)

Attribute code, name and descriptor	Reference List	Attributes compared with Radziwill and Benton [69]: New (N) Adapted (A) Retained (R)
A.21 Interaction enjoyment. The chatbot is perceived as enjoyable and engaging to operate with	[48]	A—This attribute merged two attributes ‘Entertain and/or enable the participant to enjoy the interaction’ and ‘Make tasks more fun and interesting’
A.22 Read and respond to the moods of the participant. Chatbot seems able to appropriately recognise the mood of the user from the conversation and to respond accordingly	[37, 57]	R
A.23 Sensitivity to safety and social concerns. Chatbot seems able to recognise and respond to safety or social concern and to refer a user to helpline if needed	[26, 58]	R
A.24 Meets diversity needs. Chatbot seems able to meet needs and be used by users independently from their health conditions, well-being, age, etc.	[69]	A—Original description ‘Meets neurodiverse needs such as extra response time and text interface’
A.25 Trustworthiness (general sense of trust). The chatbot is perceived as an accountable and reliable tool to enable users to achieve their goals	[2, 12, 36, 48]	A—Original description ‘Trustworthiness’
A.26 Process tracking and follow up. Chatbot seems to be able to inform and update users about their status and progresses toward the achievement of the goal	[79]	A—Original description ‘Facilitate transactions and follows up with status reports’
A.27 User’s privacy and security. Chatbot appears to be able to protect user’s privacy and make appropriate decisions on behalf of the user	[2, 79]	A—Original description ‘Protect and respect privacy’

were retained only when these were described as having a perceivable characteristic that people may use to assess and judge (a system and the experience of using that system) after they had used a CRM chatbot, to rate their experience as satisfactory or not.

In parallel, a systematic literature review was performed following PRISMA guidelines [59]. The outcomes of the review were also used to specify and add attributes to the final list. Researchers performed the initial process of review and adaptation of the list attributes and also reviewed the process and the list (see Appendix 1).

## 2.2 Results

Figure 1 reports on the PRISMA process that resulted in a final database of thirty-four literature items.

Twelve attributes from the list of Radziwill and Benton [69] were excluded because they were not relevant or not applicable for the assessment satisfaction with CRM chatbots (see Appendix 2). A revised list of 27 attributes was composed by using the remaining set of attributes from Radziwill and Benton [69] as a driver and by adding attributes in line with the new set of references (see Table 1).

## 2.3 Discussion

A total of 27 attributes was identified by extending and reviewing the previous work of Radziwill and Benton [69] for the specific

purpose of assessing user satisfaction with CRM chatbots. Using the same list mechanism proposed by Radziwill and Benton [69], these attributes could be used as a checklist in order to control the quality of the chatbot functionalities during the design phase. In order to further refine the list, a group of experts and end-users were involved in a second study to ensure that the list was developed in a robust manner.

## 3 Study 2—Attributes Selection

### 3.1 Methods

#### 3.1.1 Participants

Fifty experts and users were invited to complete an online survey based on the quality assessment attribute collection. Participants were recruited from a pool of expert designers and end-users provided by industry—the company UserBot.ai (<https://userbot.ai/>) and from the student population of the University of Twente. Twenty-nine (58%) completed the survey.

#### 3.1.2 Procedure

First, the participants were asked to complete a consent form and provide demographic data; participants indicated their role as either chatbot designers or as end-users.

Designers declared their expertise in the number of years they had worked in the field and end-users declared the amount of interaction with chatbots they had had in the last 12 months (the scale used ranged from 1 = None to 6 = Every Day). In the main part of the survey, participants rated how much they agreed with the importance of each attribute; this was accomplished using a 7-point Likert scale mechanism. Finally, participants were asked to leave comments related to (i) comprehensibility and the wordings of the attribute’s name and descriptions and (ii) missing attributes and additional aspects which they thought should be included.

### 3.1.3 Data Analysis

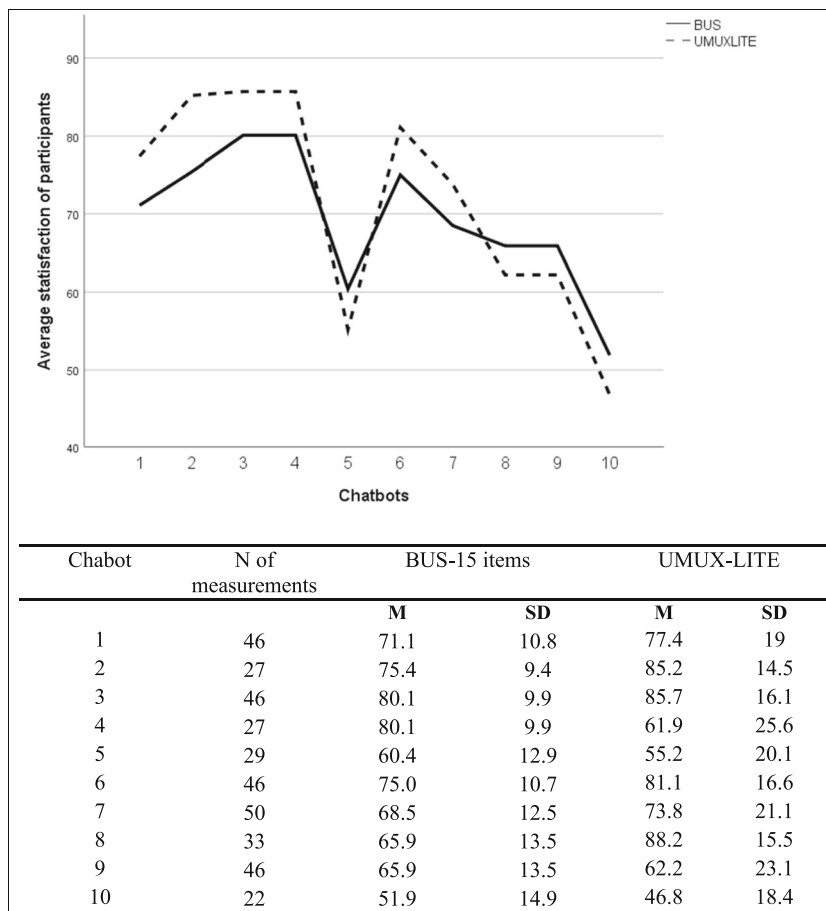
The consensus on attributes was analysed by the median scores for each factor. To be inclusive and representative of the different experiences and targeted at building a tool that could be used by end-users with different levels of expertise, we weighted the value of all of the opinions of the

stakeholders equally. Interquartile ranges (IQRs) were used to estimate the level of agreement per factor (Polisena et al., 2019). In order to be precise, only attributes with an overall median IQR between 5 (agree) and 7 (strongly) were retained in-line with Polisena et al. (2019). Moreover, agreement on the final list of attributes was estimated using Krippendorff’s Alpha with a 10,000 bootstrap resampling to estimate inter-coder reliability [35].

### 3.2 Results

Among the 29 (volunteers - 27 male, 2 female; Mean Age: 36.5; SD: 9.3) stakeholders involved in the survey: (i) eight declared themselves experts (designers or programmer) with average expertise of two years in the chatbot field. (ii) ten reported they were frequent users, having used a chatbot every day or at least once a week in the last 12 months and, (iii) eleven declared themselves novices, with minimal experience with chatbots in the last 12 months. A Consensus Analysis (Table 2) showed that only seventeen of the twenty-seven attributes that were included in the

**Fig. 2** Graphic presentation of the average score of participants’ satisfaction measured by the BUS-15 and the UMUX-LITE per chatbot. A descriptive analysis is included by reporting per chatbot for the number of participants, the mean and the standard deviation of the BUS-15 and the UMUX-LITE



revised list were considered important enough to assess ‘satisfaction’ for the different stakeholder types. The agreement among the participants was equal to 0.780, which is acceptably higher than the minimum level of .667 for the Krippendorff’s Alpha [35].

Among the attributes which were excluded, seven were related to conversational capabilities and appearance of the chatbots (A2, A9, A10, A14, A18, A19 and A22), and one

attribute was related to the sensitivity of the chatbot to recognise if people needed help or support (A24). Finally, two attributes that are usually reported in the literature as critical interactive aspects to determine the overall quality of experience with chatbots, such as ‘Personality’ (A20) and ‘Interaction enjoyment’ (A21), were not considered essential for the stakeholders of the survey to determine people’s satisfaction with CRM chatbots.

**Table 2** Agreement of the different stakeholders (Expert Designers, End-Users with a good or high level of expertise and Novices) on the importance of the attributes used to assess the quality of interaction with CRM chatbots

Attribute code	Attribute name	Expert designers’ median (IQR)	Expert users’ median (IQR)	Novice users’ median (IQR)	Overall median (IQR)	Retained (R) Excluded (E) Uncertain (U)
A1	Response time	6 (5–6.5)	6 (5–6.75)	6 (6–7)	6 (5.5–7)	R
A2	Multi-thread conversation	5 (4–6)	5 (4–6)	6 (5–7)	5 (4–6.5)	E
A3	Maxim of quantity	6 (5–6)	6 (4.25–6.75)	7 (6–7)	6 (5–7)	R
A4	Maxim of quality (perceived credibility)	7 (7–7)	6 (4.5–7)	7 (6–7)	7 (6–7)	R
A5	Maxim of manners (understandability)	7 (6–7)	6 (4.25–6)	6 (5–7)	6 (5–7)	R
A6	Maxim of relation	6 (6–7)	6 (5–6)	6 (6–7)	6 (6–7)	R
A7	Appropriate language style	6 (5.5–6)	5 (3.5–5.75)	6 (4–6)	6 (5–6)	R
A8	Reference to the service	7 (6–7)	6 (4.25–6.75)	6 (5–6)	6 (6–7)	R
A9	Visual Look	6 (4–6)	4.5 (3–5.75)	4 (3–5)	5 (3–6)	E
A10	Voice Tone	5 (4–6)	4.5 (4–6)	6 (4–6)	5 (4–5)	E
A11	Integration with the website or platform (Visibility)	6 (6–7)	5 (5–6)	7 (6–7)	6 (5.5–7)	R
A12	Graceful responses in unexpected situations	6 (6–7)	6 (6–7)	6 (3–6)	6 (5–7)	R
A13	Recognition and facilitation of users’ goal and intent	7 (6–7)	6 (5–7)	7 (6–7)	7 (6–7)	R
A14	Variation of responses	5 (4.25–5.75)	4 (3–5)	6 (5–7)	5 (4.5–6)	E
A15	Perceived Ease of Use	6 (4–6.75)	6 (5–6)	7 (6–7)	6 (6–7)	R
A16	Engage in on-the-fly problem solving	6 (5.25–7)	6 (5–6)	7 (5–7)	7 (5.5–7)	R
A17	Ability to maintain a themed discussion	6 (5.25–6.75)	5 (4–7)	6 (4–7)	6 (5–7)	R
A18	Breadth of knowledge	5 (4–5)	5 (4–7)	4 (3–4)	4 (3.5–5)	E
A19	Initiative	5 (4.25–6)	5 (3–5)	5 (3–5)	5 (4–5)	E
A20	Personality	6 (5–7)	5 (4–6)	6 (4–6)	5.50 (4–6)	E
A21	Interaction enjoyment	4.5 (2.50–5.75)	4.50 (2.50–5.25)	5 (4–6)	5 (4–6)	E
A22	Read and respond to moods of human participant	5.5 (5–6)	5 (4–6)	3 (2–5)	5 (3–5.5)	E
A23	Users’ privacy and security	7 (6–7)	7 (6–7)	6 (5–6)	6 (6–7)	R
A24	Sensitivity to safety and social concerns	6.5 (5–7)	5.5 (3.75–6.25)	5 (4–5)	5 (4–6)	E
A25	Meets neurodiverse needs	6 (5–6.75)	5 (3.75–6.25)	6 (5–7)	6 (5–6.5)	R
A26	Trustworthiness (general sense of trust)	6 (5–6)	4.50 (3.75–6.25)	6 (5–7)	6 (5–6)	R
A27	Process tracking and follow up	6.5 (6–7)	6.50 (5–7)	7 (7–7)	7 (6–7)	R

**Table 3** The revised list of attributes to assess the quality of interaction with CRM chatbots (code, name) and descriptors of new items. Attributes included in the previous list were coded A, new attributes were coded N, attributes previously excluded and re-inserted were coded R

Code	Attributes' name*
A1	Response time
A2	Maxim of quantity
A3	Maxim of quality (perceived credibility)
A4	Maxim of manners (understandability)
A5	Maxim of relation
A6	Appropriate language style
A7	Reference to the service
A8	Integration with the website or platform (visibility)
A9	Graceful responses in unexpected situations
A10	Recognition and facilitation of users' goal and intent
A11	Perceived Ease of Use
A12	Engage in on-the-fly problem solving
A13	Ability to maintain a themed discussion
A14	Users' privacy and security
A15	Meets neurodiverse needs
A16	Trustworthiness (general sense of trust)
A17	Process tracking and follow up
N18	Linguistic flexibility: Chatbot seems able to manage and adapt to different conversational styles of the end-user
N19	Ease to start a conversation: The design of the chatbot minimise the barriers and makes clear how to start a conversation
N20	Expectation setting: Chatbot makes immediately clear its capabilities and limitations without creating false expectations
R21	Interaction enjoyment
R22	Personality

\*A descriptor was added for new attributes

Participants suggested some minor changes in the wordings of the attributes to improve the readability of each attribute characterisation and highlight any omissions or perceivable errors. However, only two designers commented on potential extra attributes that were missing from the list. One designer suggested adding attributes connected to the linguistic capability of the chatbot by saying that it is vital that from the conversational point of view that a 'chatbot understands needs and mood of the users by giving precise information in as less time as possible' (D3). The other designer suggested that it is crucial that a conversational agent was set up and reframed the end-user's expectations by 'acknowledging when it doesn't have enough confidence in emitting a response' (D11).

### 3.3 Discussion

Study 2 suggested that seventeen attributes were considered the most relevant. Among the attributes that participants rated less important were attributes that were hard to judge (A24) or related to aesthetics and

conversational capabilities that could only minimally affect the overall experience of use with chatbots (*as seen in* A2, A9, A10, A14, A18, A19 and A22). Conversely, the exclusion of the attributes 'Personality' (A20) and 'Interaction enjoyment' (A21) was unexpected. These attributes seem very strongly connected to the user experience so that the personality of the chatbots and the enjoyment of interaction are often discussed as key elements to assess the adoption and use of conversational agents [11, 45, 64, 87]. The fact that CRM chatbots usually have a short-time relationship with end-users could have led to the exclusion of these attributes [28], whereas attributes such as A20 and A21 could be more seen more critical in judging conversational agents which are meant for long-term interactional exchanges. Attributes A20 and A21 are excluded from this study that is focused on CRM chatbots, but these could be used to assess satisfaction with conversational agents. The results also suggested that other potentially essential attributes could be included in the list concerning linguistic capabilities and user expectations. These suggestions seem in line with indications of



**Table 4** Revised list of key attributes from Study 1 (code and names). These attributes were listed as essential aspects to assess the quality of interaction with CRM chatbots after the focus group. The participants’ agreement on the importance of each attribute, and indications emerged during the focus group were used to decide whether to retain (R), change (C) or merge (M) attributes. The rationale behind the decision-making is reported together with the final list of attributes (name and amended descriptions)

Attribute code	Participants’ agreement	Retained (R) Changed (C) Merged (M) Excluded (E)	Rationale	Final attribute name and new/modified descriptor
A3 A16	100% 81%	M	These two attributes were often confused by end-users. Perceived credibility was perceived as an attribute that already covers trustworthiness	Perceived conversational credibility The chatbot responds in a credible and informative way without adding too much information
A1 A12 A17	100% 73% 80%	M	Participants find hard to understand the difference between A1, A12 and A17 and suggested to reword these into a new attribute that measures how quick chatbot is able to answer to a request	Speed of answer The chatbot is perceived as able to respond to requests and solve issues in a timely manner
A4 A6	100% 80%	M	Participants mainly interpret attribute A4 associated with A6. By suggesting that until a chatbot is not rude what it really counts from an end-user perspective is the tool is able to understand input and to provide understandable output	Understandability and politeness The chatbot seems able to understand input and convey correct statements and answers without ambiguity and with acceptable manners
A2	100%	R	-	Maxim of quantity
A7	94%	R	-	Reference to service
A13	88%	R	-	Ability to maintain a themed discussion
A10	93%	R	-	Recognition and facilitation of users’ goal and intent
A14	93%	C	Participants suggest making clear that the attribute is about the perception of how the chatbot enables privacy	Perceived privacy and security
A9	69%	C	Participants had difficulty to understand the relevance of attribute because of the wording and suggested to reword it	Resilience to failure. Chatbot seems able to find ways to respond appropriately even when it encounters situations or arguments it is not equipped to handle
N19	80%	R	-	Ease to start a conversation
A8	93%	C	Participants suggested changing the name and the description to reflect that possibility to get access to the tool and its functions	Access to Chatbot Functions and location of the chatbot on the screen are visible and accessible
N20	100%	R	-	Expectation setting
A5	87%	C	Participants suggested to avoid jargon and make clear the name of the attribute and its descriptor	Relevance of information The chatbot provides relevant and appropriate information/answer to people at each stage to make them closer to their goal
N18	94%	C	Participants suggested making clear that this attribute is related to the capacity of chatbot to answer despite the different styles of input provided by end-users	Flexibility and communication effort Chatbot seems able to manage and adapt to different conversational styles of the end-users minimising conversational efforts for the end-user
R21 R22	50% 50%	E E	All participants agreed that these two aspects are not really important from the satisfaction point of view, despite these could enhance user experience. This result was consistent with the outcomes of Study 1	-
A11	100%	E	Despite all participants agree that on its own A11 (ease of use) is an important attribute, this attribute was considered too vague	-

Zamora [87] who reported that aspects such as the capabilities of chatbots to accommodate to different conversational style, and the ability to make it easy for

end-users to start a conversation and achieve relevant results is essential to boost the experience of interacting with conversational agents.

**Table 5** Loading of attributes and items and excluded items

Attribute	Items	Factor loadings					Excluded
		1	2	3	4	5	
1. Ease to start a conversation	1	1					
	2	.999					
	3	.999					
2. Access to chatbot	4	.999					
	5	1					
	6	1					
3. Expectation setting	7		.999				
	8		.983				
	9		.772				
4. Flexibility and communication effort	10	.979					
	11	.993					
	12		.907				
5. Ability to maintain a themed discussion	13		.923				
	14		.993				
	15		.998				
6. Reference to the service	16			.980			
	17		.998				
	18		.971				
7. Users' privacy and security	19				.981		
	20				.981		
	21				.981		
8. Recognition and facilitation of users' goal and intent	22			.96			
	23			.98			
	24			.98			
9. Relevance of information	25			.98			
	26			.98			
	27			.98			
10. Maxim of quantity	28			.979			
	29			.98			
	30			.98			
11. Resilience to failure	31		.88				
	32		.99				
	33		.998				
12. Understandability and politeness	34		.816				
	35		.999				
	36		.999				
13. Perceived conversational credibility	37			.981			
	38		.517				E
	39			.976			
14. Speed of answer	40					.989	
	41					.989	
	42					.989	

## 4 Study 3—Revision of Attributes, Item Generation and Focus Groups

### 4.1 Methods

In line with the recommendations provided by participants of the previous study and supported by the research literature, three more attributes were added to the list:

i) Linguistic flexibility. This attribute refers to the perceived capabilities of the chatbot to manage and adapt to the different conversational styles by avoiding, for instance, that end-users should rephrase input in different ways to get answers from the conversational agent [15, 47, 87];

ii) Easiness to start a conversation. This attribute refers to the affordances provided by the design of the chatbot to make it easy for an end-user to understand how to initiate a conversation [10, 87].

iii) Expectations setting. This attribute refers to the ability of a chatbot to make clear its capabilities and not to create false expectations in the end-users [5, 32, 45, 87].

Moreover, attributes that were unexpectedly excluded in study 1 (Personality and Enjoyment) were re-inserted in the list to double-check their importance further. Therefore, the list of attributes for Study 2 was composed of 22 elements (see Table 3).

**Table 6** Reliability analysis of the items: Initial reliability estimated per each factor, reliability expected when dropping items, retained (R) items and the final alpha of each factor after dropping items

Factors	Alpha	Attributes	items	Reliability if an item is dropped	Retained	Final alpha			
1	.77	Ease to start a conversation	1	.72		0.87			
			2	.82					
			3	.82					
		Access to chatbot	4	.82					
			5	.80	R				
			6	.81	R				
			10	.86					
		Flexibility and communication effort	11	.86					
			11	.86					
		2	.85	Expectation setting	7		.5	R	0.74
					8		.38	R	
Flexibility and communication effort	9			.48					
	12			.35					
	13			.29	R				
Ability to maintain a themed discussion	14			.4	R				
	15			.61					
	17			.45					
Reference to the service	18			.61	R				
	31			.66	R				
Resilience to failure	32			.55					
	33			.58					
	34			.5					
	35			.38					
	36			.48	R				
	16			.79					
	22			.64					
3	.95	Reference to the service	23	.81		.86			
			24	.77	R				
		Recognition and facilitation of users' goal and intent	25	.85					
			26	.82					
			27	.89					
		Relevance of information	28	.65					
			29	.75	R				
			30	.75	R				
		Maxim of quantity	37	.79	R				
			39	.79					
39	.79								
4	.87	Users' privacy and security	19	.78		-			
			20	.68	R				
			21	.82					
			40	.93					
5	.93	Speed of answer	41	.84	R	-			
			42	.91					
			42	.91					

A panel of four experts on interaction (three junior experts, external to the previous phase and one of the authors) proposed for each of the attribute a list of three items to create a questionnaire. Across the groups, similar items were merged, and the wording of each item was discussed in multiple sessions. However, only 21 of the attributes in the list were used. Attribute A15 was excluded from this exploratory study because it was not possible to recruit people with disability for the panel or the focus group. In this sense, by endorsing the motto ‘Nothing About Us Without Us!’ [9], the authors of the present work decided to postpone and adapt in future studies the scale by including people with disabilities. Agreement amongst panel members was reached on 61 of the 63 items generated (Appendix 3); therefore, two items were excluded from the study.

Focus group sessions were performed in order to revise the wording of the items and to inspect whether the connection between items and attributes was understandable for potential end-users.

**4.1.1 Participants**

A total of 16 volunteers (8 female & 8 male, Age M.= 22.1, SD = 2.84) participated in the focus group sessions. Participants were randomly assigned to a focus group session with a maximum of five people.

**4.1.2 Material**

During the focus group, the list of attributes (seen in Table 3) and the list of items (Appendix 3) were reviewed by

**Table 7** Correlation between UMUX-LITE, the five factors of the BUS-15 and the overall BUS-15 scale

Chatbot	UMUX-LITE					BUS-15
	F1	F2	F3	F4	F5	
1		.762**	.737**			.707**
2	.676**	.551**	.586**		.494**	.749**
3		.742**	.771**			.610**
4	.515**	.777**	.817**	.417*	.450*	.817**
5	.484**	.634**	.654**		.431*	.670**
6		.717**	.779**		.369*	.689**
7	.517**	.788**	.736**	.283*		.751**
8		.704**	.825**			.652**
9		.816**	.825**			.705**
10	.558**	.813**	.860**			.707**

\*\*significant at the 0.01 level (2-tailed)\*\*

\* significant at the 0.05 level (2-tailed)\*

participants. Consent and demographic data were obtained by Qualtrics (Appendix 4). Moreover, a demonstration to exemplify the interaction with service chatbots was given by using an actual bot; the Finnair Messenger was used: (<https://www.messenger.com/t/Finnair>). The Finnair Messenger represented a real-world example of a CRM chatbot which is integrated into a social media platform. Each session of the focus group was both audio and video recorded to facilitate and support the analysis to provide reliable data.

#### 4.1.3 Procedure

Each participant was asked to fill in a consent form and demographic questionnaire. A definition of CRM chatbots and conversational agents was given to and discussed with the group. During the demonstration, the moderator operated the Finnair chatbot while asking the participants to offer input. Following the demonstration, the moderator asked participants to reflect and discuss the positive and negative aspects of interaction with chatbots. At the end of the discussion, participants were asked to:

- i. Review the list of attributes. Each participant was provided with the list of attributes, and they were asked to discuss each attribute in terms of relevance to assess their satisfaction in the use of a CRM chatbot and to review the clarity of the attributes' descriptors.
- ii. Review the list of items: Each participant was provided with a list and asked to read the list of items to comment about the clarity of the wordings, and they were also asked to express verbally any unclear association between items and attributes. It was explained to participants that

an item could be matched to several attributes or none if they thought this was the case.

#### 4.1.4 Data Analysis

The panel reviewed video recordings and notes of the focus group session to:

- i. Change or adapt the list of attributes: Positive written indication and verbal comments of the participants about the importance of each attribute, and the comprehensibility of its descriptor in the list was used to assign the value 1 to indicate that the attribute was comprehensible and considered necessary by a participant. Conversely doubt about the attribute, its descriptor and its importance to assess satisfaction was coded as '0'. Positive responses were used to estimate the level of agreement on the relevance of each attribute to assess user satisfaction during the use of service chatbots.
- ii. Change or adapt the list of items: Comments of participants about ambiguity in the item's wording, typos, or unclear association between items and attributes were noted during the focus group and analysed post-session using video recordings.

## 4.2 Results

As reported in Table 4, the initial list of 21 attributes was reduced to a list of 14 main attributes. Attributes R21 and R22 (Enjoyment and Personality) were excluded. As was the case in Study 1, these attributes were not considered as an essential factor in assessing the satisfaction with CRM chatbots. The attribute A11 (ease of use) despite being considered important as a factor was described by 15 out of 16 participants as too vague. Participants suggested that 'Ease of use' was already covered by other attributes and that each person may have a different idea of what 'easy to use' entails. Similarly, A11 was also excluded from the list. Moreover, the description of five attributes was slightly adjusted, to avoid ambiguity, concerning the feedback data from the participants.

Participants also suggested merging the following attributes:

- Attributes A3 (Maxim of quality) and A16 (Trustworthiness) were often confused by participants who reported that to judge the trustworthiness of a chatbot; they will rely on its ability to act and respond credibly. In agreement with participants, we only retained items of A3 (see Appendix 3) to measure a new attribute that we named: 'Perceived conversational credibility'.

- Attributes A1 (Response time), A12 (Engage in on-the-fly problem solving) and A17 (Process tracking and follow up) were considered by participants all attributes related to the ability to answer in a quick way to the request of end-users. In agreement with participants, we only retained items of A1 (see Appendix 3) to measure the new attribute ‘Speed of answer’.
- Attributes A4 (Maxim of manners) and A6 (Appropriate language style) were both considered associated. Therefore, items of A4 (see Appendix 3) were retained to measure the new attribute ‘Understandability and politeness’.

Regarding the quality of the items wording, no major request for changes was outlined, despite some typos were highlighted by participants. Therefore, all the proposed items were corrected and retained for further testing. In tune with the indication from the focus group, the preliminary version of the BUS was composed of 42 items associated with 14 attributes (see Appendix 5).

### 4.3 Discussion

Participants of Study 3, in line with results of Study 2, suggested excluding the attributes ‘Interaction Enjoyment’ and ‘Personality’. This seems to confirm that these attributes are considered less important than others by end-users to assess the satisfaction with CRM chatbots, or as earlier mentioned, too generic and addressed by other factors in the scale. However, as stated earlier, these two attributes should be considered and employed when dealing with chatbots for long-term interaction/relationship-based interaction. It is also worth discussing the exclusion from the attribute list ‘ease of use’. The overall perspective of the participants was that ‘ease of use’ could not be fully represented by one attributional factor, but that the ability to judge ‘ease of use’ with a CRM chatbot is something that could emerge by considering a related set of interactive and conversational factors during the exchanges with chatbots. Participants in the focus groups also considered those attributes and items that could be concretely perceived and observed during the interaction as relevant. Participants agreed that from an end-user perspective:

- i) It is easier to assess ‘trust’ in a CRM chatbot interaction by assessing the bot’s capacity to provide information and helping to attain a goal (i.e. the credibility of information) instead of by assessing trustworthiness as a general and unspecified sense of trust. Assessing trustworthiness could require a different set of items more in line with trust and technology acceptance theory [55].
- ii) The ability of chatbots to provide speedy (and accurate) answers to their request was considered easier to assess,

than its capacity to solve emerging issues or its ability to inform them about their progress toward the achievement of the goal.

- iii) It was more comfortable and more relevant to assess the capability of chatbots to understand and be understandable than its ability to use an appropriate style of language.

The list of 14 attributes resulted from the analysis is reported in Appendix 6 as a checklist to assess the quality of chatbots (BOT-Check). BOT-Check could be used to enable designers to control quality during the development of CRM chatbots, i.e. agents for short-term interaction. Moreover, by adding three other attributes to the list that were excluded from the present work as previously discussed, such as ‘Interaction Enjoyment’ and ‘Personality’ and ‘Meets neurodiverse needs’ designers, could aim to assess long-term conversational agents more inclusively.

## 5 Study 4. Psychometric Exploration of the BUS

A test was performed with participants interacting with multiple chatbots (five out of ten) to explore the psychometric properties of the scale (BUS-42) and to reduce the number of items systematically.

### 5.1 Methods

#### 5.1.1 Participants and Measures

A total of 480 questionnaires were collected from a sample of 96 volunteers (22 Female, 74 Male Age M: 23.7, SD: 4.8). Eight percent (385) of the questionnaires were entirely or correctly completed.

#### 5.1.2 Material

Ten chatbots were used in this pilot study which used the scale; each one of these was associated with an information retrieval task (Appendix 7). Qualtrics was used to collect information relating to demographics (see Appendix 4), to present the tasks to be accomplished and to collect feedback after the use of each chatbot using the 42-item BUS and a UMUX-LITE [53]. Each item of the BUS was presented as a statement to the participants, and they were asked to assess their agreement with each statement on a five-point Likert scale from 1 (‘Strongly Disagree’) to 5 (‘Strongly Agree’). A five-point Likert scale version of the UMUX-LITE was used in line with the recommendations of Sauro [73] and Lewis [51].

### 5.1.3 Procedure

Participants were tested in a dedicated room. Consent and demographic information were acquired, and participants were asked to interact randomly with five of the ten chatbots available (see the list of chatbots and tasks, in Appendix 7) to achieve a goal; this was presented as an information-retrieval task. After the interaction with each chatbot, if the participants achieved the task or not, they were required to fill the 42-item BUS and the UMUX-LITE, and they then had a 10-min break. Each participant used the same computer and monitor for the test. As some of the data were collected before and during the pandemic crisis due to COVID19, 60% of the data were collected in presence, and 40% of the data were collected by in-presence remote testing mediated by video calling systems with the same procedure of the in-presence collection.

### 5.1.4 Data Analysis

The 385 questionnaires were used to perform a 50,000 iterations Bayesian Exploratory Factor Analysis (BEFA, [16]) with R package ‘BayesFM’ [66]. ‘Psych’ R package was used to perform a parallel analysis [70]. Multiple BEFA were performed as defined by Conti et al. [16] suggested that BEFA is an iterative approach which reduces items and analyses factor-loading. Bayesian approaches of factorial analysis are considered more reliable compared with classic approaches [39]. Reliability analysis was conducted individually for each latent factor using the alpha function from the R package ‘psych’ [70]. This analysis was used to drop items and improve internal consistency systematically. Finally, participants’ answers (per chatbot) were used to perform descriptive and Pearson correlation analyses to explore the relationship among the final version of the BUS (and its factors) and the UMUX-LITE.

## 5.2 Results

### 5.2.1 Bayesian Exploratory Factor Analysis

A parallel analysis suggested a structure with five components. The BEFA analysis confirmed the structure with five factors (35%, Metropolis-Hastings acceptance rate = 0.996). In tune with DeVellis [21], we only retained the items with loading over 0.7 (Table 5).

### 5.2.2 Internal Consistency

By aiming at reducing the number of items and concurrently maintaining a level of reliability above .7 for each factor, multiple iterations of reliability analysis were performed by dropping items iteratively until a satisfactory solution was identified. Coherence between the attributes

associated in each factor was also considered to exclude or retain an item.

As reported in Table 6 the final questionnaire was reduced to 15 items (BUS-15, see Appendix 8) as follows:

- Factor 1, initially composed of 8 items ( $\alpha=.77$ ) was reduced to 2 items ( $\alpha=.87$ ). This factor was named ‘Perceived accessibility to chatbot functions’ intended as the design of the chatbot to enable users to start a conversation and to achieve their goal.
- Factor 2, initially composed of 14 items ( $\alpha=.85$ ) was reduced to 7 items ( $\alpha=.74$ ). This factor was named ‘Perceived quality of chatbot functions’ intended as the ability of the chatbot to communicate its functions and use the information available on the screen to drive people’s interaction in a polite way and in line with end-user expectations.
- Factor 3 initially composed of 12 items ( $\alpha=.95$ ) was reduced to 4 items ( $\alpha=.86$ ) after repeated dropping of items. This factor was named ‘Perceived quality of conversation and information provided’ intended as the perceived ability of a chatbot to engage in a conversation adequately.
- Factor 4 initially composed of 3 items ( $\alpha=.87$ ) was reduced to one item regarding privacy and security of interaction exchange. This factor was named ‘Perceived privacy and security’ intended as the perceived ability of the chatbot to enable people to achieve their goal.
- Factor 5 composed of 3 items ( $\alpha=.92$ ) was reduced to one item concerning the response waiting time. Therefore, this factor was named ‘Time response’.

When all the items included in the BUS-15 are considered, the overall alpha was equal to .87.

### 5.2.3 Relationship between BUS-15 and UMUX-LITE

Figure 2 reports the average reaction to the different chatbots under assessment measured by UMUX-LITE and by BUS-15. A total of 13 participants only partially completed the UMUX-LITE. Therefore, this analysis was performed on 372 valid measurements. The satisfaction measured by BUS-15 spanned from a min. of 51.9% to a max. of 80.1% compared with the UMUX-LITE results that spanned from a min. of 46.8% to a max. of 88.1%.

Table 7 suggests that by looking at the results per chatbot, the UMUX-LITE and the overall scale of BUS-15 strongly correlate, however, the five factors of the BUS-15 seem to provide a broader perspective and capture aspects not considered by UMUX-LITE. Specifically, two factors of BUS-15, namely, ‘Perceived quality of chatbot functions’ (F2), ‘perceived quality of conversation and information provided’

(F3) consistently correlate with the average items of UMUX-LITE. In comparison, three factors, namely, ‘Perceived accessibility to chatbot functions’ (F1), ‘Perceived privacy and security’ (F4) and ‘time response’ (F5) seem to have mild correlation or to not correlate with UMUX-LITE on several occasions.

### 5.3 Discussion of Study 4

The exploratory analysis we performed suggested that with 15 items, the BUS could reliably enable end-users to express their perception about their experience with a chatbot. The overall scale of BUS seems to strongly correlate with the ultra-short and unidimensional standardised measure of satisfaction proposed by the UMUX-LITE. However, BUS-15, with its five factors, would still enable the assessment of differences in people’s perspectives by considering aspects such as accessibility to the chatbots’ functions, time to response and privacy. Factors not usually considered as ‘classic’ measurements of satisfaction would be developed for non-conversational tools.

The current version of the BUS-15 (Appendix 8) should be considered an initial step into a somewhat uncharted domain, i.e. the assessment of satisfaction with conversational agents. This scale could be applied to practical use or used to get comparable data among/across chatbot-based tools/systems or during cycles of design and redesign; however, the results cannot be yet considered conclusive and further studies are needed to extend and revise the construct and to validate the scale fully. Conversely, BOT-Check (Appendix 6) could be used by designers as a tool to ensure quality in the design and functioning of chatbots before the testing with end-users. This checklist should be considered complementary to the use of BUS-15.

## 6 Conclusion

The advantage of having a reliable scale to test people’s perception of the quality of interaction with conversational agents is that such a tool may enable (i) potential end-users to express their level of satisfaction in a consistent and replicable way, (ii) designers and evaluators to develop benchmarks to compare their results by modelling the different end-users and their need during the formative and summative phase of product assessment. Currently, BOT-Check could be considered a ready to use diagnostic tool to control how much a chatbot interacts with people in line with guidelines and principles of quality design for conversational agents e.g. heuristic inspection. Conversely, BUS-15 currently cannot be used as an off-the-shelf product for user research and usability tests. Although we included a reasonable number of chatbots widely used by customers, further validation studies are needed with a larger number of chatbots and a diverse range of

participants to ensure the reliability of the construct and to streamline the current version of BUS. During the testing as part of the exploratory analysis of the BUS, some tools were closed for proprietary reasons or temporarily suspended due to COVID19, e.g. <https://www.ato.gov.au/>. This was not an issue, as we were able to collect data to perform the analysis; however, it is representative of the volatile nature of the market for CRM chatbots. The threats to the validity of the present study should also be considered before using BUS-15. As we stated earlier, a more diverse range of people (age, gender and ability) are needed to use the system in future iterations; in this study mainly young participants with age below 35 years old were involved in focus groups and in the pilot of the scale. A more systematic analysis of people should be performed in future works to capture the perspective of different potential end-users better. Concurrently, as we reported above, the present version of the construct did not include the perspective of people with disabilities, and future research and evaluations should plan for this.

Despite the limitations, the present work provides a new list of attributes specifically developed to measure satisfaction with CRM chatbots and a preliminary tool for assessment. We invite practitioners and researchers who want to contribute to the development of this tool to use BUS, together with other tools, as a way to get insights about the needs and the point of view of end-users about the interaction with a chatbot.

Conversational agents are creating an interactional paradigm shift and a range of new research and design opportunities in the field of HCI [27]; nevertheless, the quality of interaction with these tools can only be ensured by defining reliable criteria and assessment tools that can ensure comparability and support a satisfactory exchange between people and this evolving type of intelligent technology.

**Acknowledgement** We would like to thank the company Userbot.AI and its team, which supported the recruitment of experts and participants to build the construct behind the scale. Moreover, we would like to thank students Lisa Waldera, Nina Böcker, Alexander Dehmel, Steffen Neumeister, for their help in gathering the data.

**Funding** We are grateful to the UKRI project Not-Equal, funded by EPSRC through the Digital Economy Theme (EP/R044929/1), for partially funding this research through the call for collaboration project MiniCoDe – Minimise algorithmic bias in Collaborative Decision Making with Design Fiction. Dr Alan Chamberlain’s part in this work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T51729X/1] projects RCUK Catapult Researchers in Residence award Digital - Disruptive Beats - Music - AI - Creativity - Composition and Performance, [grant number EP/V00784X/1] UKRI Trustworthy Autonomous Systems Hub and [grant number EP/S035362/1] PETRAS 2.

Dr Simone Borsci is also affiliated to the National Institute for Health Research, London IVD Co-operative, Faculty of Medicine, Department of Surgery & Cancer, Imperial College, London, UK, and to the Schools of Creative Arts, University of Hertfordshire, Hertfordshire, U.K.

## Appendix 1 Systematic literature review plan (PRISMA Checklist)

### This literature review's contribution to existing research:

fill the gap in the literature by defining a list of criteria of quality for the interaction with a chatbot for information retrieval tasks.

The focus	Peer-review articles and conference papers that include findings and theories of quality of interaction with chatbots and articles that will include assessment methods for interaction quality with chatbots
The goal	Integrate and generalise previous findings and propose a list of the key factors that affect interaction with a chatbot
Perspective	The language of the literature review will be neutral
Coverage	The review will only cover central or pivotal literature
Organisation	The review will be organised around the propositions in a research rationale
Audience	<i>Primary</i> —Reviewers of the work GM, SB, DB <i>Secondary</i> —Co-authors, other scientists, experts that were included in the research
Methodology	This literature review is qualitative and will follow the phenomenological method of the literature review
Inclusion criteria	<ul style="list-style-type: none"> <li>• Studies that mention chatbots or conversational interfaces/agents in their Title, Abstract or Keywords</li> <li>• Studies that include findings and theories on factors/aspects/attributes that can potentially contribute to the perceived interaction quality with chatbots</li> <li>• Studies inform about criteria used during the assessment of interaction with chatbots</li> </ul> Database search: <ul style="list-style-type: none"> <li>• Studies from the past 10 years</li> </ul>
Inclusion criteria	<ul style="list-style-type: none"> <li>• Items that talk about technical aspects of the chatbots</li> <li>• Items about virtual assistants and conversational agents for general purposes, i.e. not for service or information retrieval</li> <li>• Items that did not inform about interaction characteristics</li> <li>• Items that were not able to explain or clearly define attributes</li> </ul>
Search inquiry—Scopus	( TITLE-ABS-KEY ( chatbot* OR 'conversational agents*' OR 'conversational interface*' ) AND TITLE-ABS-KEY ( interact* ) AND TITLE-ABS-KEY ( satisf* OR quali* ) AND NOT TITLE-ABS-KEY ( 'virtual assistant' OR voice ) ) AND PUBYEAR > 2009 AND ( LIMIT-TO ( SRCTYPE , 'p' ) OR LIMIT-TO ( SRCTYPE , 'j' ) ) AND ( LIMIT-TO ( LANGUAGE , 'English' ) ) AND ( LIMIT-TO ( DOCTYPE , 'cp' ) OR LIMIT-TO ( DOCTYPE , 'ar' ) )
Search inquiry—Web of Science	You searched for: TOPIC: (Chatbot* OR 'conversational agent*' OR 'conversational interfaces') AND TOPIC: (Interact* AND TOPIC: (Satisf* or qual*)) Refined by: LANGUAGES: ( ENGLISH OR PORTUGUESE ) AND PUBLICATION YEARS: ( 2018 OR 2014 OR 2010 OR 2017 OR 2013 OR 2016 OR 2012 OR 2015 OR 2011 ) AND DOCUMENT TYPES: ( PROCEEDINGS PAPER OR ARTICLE ) Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.
Tools	Prisma Flow diagram, PRISMA 2009 Checklist ( <a href="http://prisma-statement.org/">http://prisma-statement.org/</a> )



## Appendix 2 Attributes of the previous review excluded

Attribute from Radziwill and Benton	Reason for excluding
1. Accurate speech synthesis	Not all chatbots have speech synthesis; therefore we excluded this attribute
2. Passes the Turing test	These attributes are mainly related to design aspects that cannot be evaluated by end-users outside of specific experimental settings
3. Does not have to pass the Turing test	
4. Transparent to inspection discloses its chatbot identity	
5. Include errors to increase realism	These are attributes that designers should aim to include in their chatbots but it is not clear how end-users can perceive or become aware of these aspects during the interaction
6. Ethics and cultural knowledge of users	
7. Awareness of trends and social context	
8. Respect, inclusion and preservation of dignity	
9. Non-deception	This is not an attribute of the chatbot but a measure of the interaction exchange
10. Convincing, satisfying and natural interaction	
11. Exude warmth and authenticity	This could be connected somehow to trust and enjoyment but it is not clear how warmth and authenticity are connected and how to measure these
12. Convincing, satisfying and natural interaction	This is not an attribute but the result of the interaction

## Appendix 3 Initial items pool

Attribute code	Attribute name	Item 1	Item2	Item3
A1	Response time	The time of the response was reasonable	My waiting time for a response from the chatbot is short	The chatbot is quick to respond
A2	Maxim of quantity	The amount of received information was neither too much nor too less	The chatbot gives me the appropriate amount of information	The chatbot only gives me the information I need.
A3	Maxim of quality (perceived credibility)	I feel like the chatbot’s responses were accurate	I believe that the chatbot only states reliable information	It appeared that the chatbot provided accurate and reliable information
A4	Maxim of manners (understandability)	I found the chatbot’s responses clear	The chatbot only states understandable answers	The chatbot’s responses were easy to understand
A5	Maxim of relation	The chatbot gave relevant information during the whole conversation	The chatbot is good at providing me with a helpful response to any point of the process	The chatbot provided relevant information as and when I needed it
A6	Appropriate language style	The style of language used by the chatbot felt appropriate	The chatbot is answering with the right amount of formality	The chatbot communicates with an appropriate language style.
A7	Reference to the service	The chatbot guided me to the relevant service	The chatbot is using hyperlinks to guide me to my goal	The chatbot offers me relevant functions to achieve the goal*
A8	Integration with the website or platform (Visibility)	The chatbot was easy to access	The chatbot function was easily detectable	It was easy to find the chatbot

A9	Graceful responses in unexpected situations	The chatbot could handle situations in which the line of conversation was not clear	The chatbot explained gracefully that it could not help me	When the chatbot encountered a problem, it responded appropriately
A10	Recognition and facilitation of users' goal and intent	I felt that my intentions were understood by the chatbot	The chatbot was able to guide me towards my goal	I find that the chatbot understands what I want and helps me achieve my goal
A11	Perceived Ease of Use	The interaction with the chatbot felt easy	I had to put in only minimal effort to use the chatbot.	I find the chatbot easy to use
A12	Engage in on-the-fly problem solving	The chatbot solved my problems instantly	The chatbot is able to answer any questions within a few seconds	The chatbot was able to engage with any request in an acceptable time frame*
A13	Ability to maintain a themed discussion	The interaction with the chatbot felt like an ongoing conversation	The chatbot was able to keep track of context	The chatbot maintains a relevant conversation
A14	Users' privacy and security	The interaction with the chatbot felt secure in terms of privacy	I believe the chatbot is informing me of any possible privacy issues	I believe that this chatbot maintains my privacy
A15**	Meets neurodiverse needs	--	--	--
A16	Trustworthiness (general sense of trust)	I felt that I could trust the chatbot	The chatbot reassures me that I can trust this technology	I trust this chatbot
A17	Process tracking and follow up	I was adequately updated about my task progress	The chatbot is giving me feedback about the status of my request	The chatbot keeps me aware of what it is doing
N18	Linguistic Flexibility	I had to rephrase my input multiple times for the chatbot to be able to help me	I had to pay special attention regarding my phrasing when communicating with the chatbot	It is easy to tell the chatbot what I would like it to do
N19	Ease to start a conversation	It was clear how to start a conversation with the chatbot	It was easy for me to understand how to start the interaction with the chatbot	I find it easy to start a conversation with the chatbot
N20	Expectation setting	Communicating with the chatbot was clear	I was immediately aware of what information the chatbot can give me	It is clear to me what the chatbot can do
R21	Enjoyment	I enjoyed interacting with the chatbot	The chatbot made it fun to research the information	The chatbot was fun to interact with
R22	Personality	The chatbot seemed like a human with its own personality	The chatbot communicated in a pleasant way with me	I found the chatbot to be likeable

\*A minimum agreement of 3 out of 4 members of the panel was not reached for this item and it was excluded

\*\*Attribute excluded from the present study because people with disability were not included in the panel nor in the focus group session

## Appendix 4. Demographic questionnaire

1. Gender
2. Age
3. Nationality
4. Education level
5. Type of education (field domain)
6. How familiar are you with chatbots and/or other conversational interfaces? (5 point Likert: Extremely familiar – Not familiar at all)
7. Have you used a chatbot or a conversational interface before? (5 point Likert: Definitely yes – Definitely no)
8. How often do you use chatbots weekly? (5 point Likert: Daily – Never)

## Appendix 5. Bot Usability Scale (42 Items)

Attributes	Items order
1. 1. 1. Ease to start a conversation	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. 3. I find it easy to start a conversation with the chatbot.
2. 1. 1. Access to chatbot	4. The chatbot was easy to access. 5. The chatbot function was easily detectable. 6. It was easy to find the chatbot.
3. 1. 1. Expectation setting	7. Communicating with the chatbot was clear. 8. I was immediately made aware of what information the chatbot can give me. 9. It is clear to me early on about what the chatbot can do.
4. 1. 1. Flexibility and communication effort	10. I had to rephrase my input multiple times for the chatbot to be able to help me. 11. I had to pay special attention regarding my phrasing when communicating with the chatbot. 12. It was easy to tell the chatbot what I would like it to do.
5. 1. 1. Ability to maintain a themed discussion	13. The interaction with the chatbot felt like an ongoing conversation. 14. The chatbot was able to keep track of context. 15. The chatbot maintained a relevant conversation.
6. 1. 1. Reference to the service	16. The chatbot guided me to the relevant service. 17. The chatbot is using hyperlinks to guide me to my goal. 18. The chatbot was able to make references to the website or service when appropriate.
7. 1. 1. Users' privacy and security	19. The interaction with the chatbot felt secure in terms of privacy. 20. I believe the chatbot informs me of any possible privacy issues. 21. I believe that this chatbot maintains my privacy.
8. 1. 1. Recognition and facilitation of users' goal and intent	22. I felt that my intentions were understood by the chatbot. 23. The chatbot was able to guide me to my goal. 24. I find that the chatbot understands what I want and helps me achieve my goal.
9. 1. 1. Relevance of information	25. The chatbot gave relevant information during the whole conversation. 26. The chatbot is good at providing me with a helpful response at any point of the process. 27. The chatbot provided relevant information as and when I needed it.
10. 1. 1. Maxim of quantity	28. The amount of received information was neither too much nor too less. 29. The chatbot gives me the appropriate amount of information. 30. The chatbot only gives me the information I need.
11. 1. 1. Resilience to failure	31. The chatbot could handle situations in which the line of conversation was not clear. 32. The chatbot explained gracefully when it could not help me. 33. When the chatbot encountered a problem, it responded appropriately.
12. 1. 1. Understandability and politeness	34. I found the chatbot's responses clear. 35. The chatbot only states understandable answers. 36. The chatbot's responses were easy to understand.
13. 1. 1. Perceived conversational credibility	37. I feel like the chatbot's responses were accurate. 38. I believe that the chatbot only states reliable information. 39. It appeared that the chatbot provided accurate and reliable information.
14. Speed of answer	40. The time of the response was reasonable. 41. My waiting time for a response from the chatbot was short. 42. The chatbot is quick to respond.

## Appendix 6. Bot Checklist (BOT-Check)

This checklist is built in a modular way to be used by designers to assess the quality of chatbots for short and long term interaction.

Modules	Attributes
1. when assessing the quality of CRM agents (short term interaction)	<p>1. Ease to start a conversation Chatbot seems able to find ways to respond appropriately even when it encounters situations or arguments it is not equipped to handle</p> <p>2. Access to chatbot Functions and location of the chatbot on the screen are visible and accessible</p> <p>3. Expectation setting Ability of chatbot to make clear its capabilities and to not create false expectations in the end-users</p> <p>4. Flexibility and communication effort chatbot seems able to manage and adapt to different conversational styles of the end-users minimising conversational efforts for the end-user.</p> <p>5. Ability to maintain a themed discussion Chatbot maintains a conversational theme once introduced and keep track of the context to understand the user's utterances</p> <p>6. Reference to the service Chatbot seems designed to use the environment (information, options, buttons on-screen, etc.) to guide the user towards its goal</p> <p>7. Users' privacy and security Chatbot appears to be able to protect user's privacy and make appropriate decisions on behalf of the user.</p> <p>8. Recognition and facilitation of users' goal and intent Chatbot seems able to recognise the user's intent and guide the user to its goals.</p> <p>9. Relevance of information The chatbot provides relevant and appropriate information/answer to people at each stage to make them closer to their goal.</p> <p>10. Maxim of quantity The chatbot responds in an informative way without adding too much information.</p> <p>11. Resilience to failure Chatbot seems able to find ways to respond appropriately even when it encounters situations or arguments it is not equipped to handle</p> <p>12. Understandability and politeness The chatbot seems able to understand input and convey correct statements and answers without ambiguity and with acceptable manners</p> <p>13. Perceived conversational credibility The chatbot responds in a credible and informative way without adding too much information.</p> <p>14. Speed of answer The chatbot is perceived as able to respond to requests and solve issues in a timely manner</p> <p>15. Meet the neurodiverse needs Chatbot seems able to meet needs and be used by users independently from their health conditions, well-being, age, etc.</p>
2. to add when assessing agents for long-term interaction	<p>16. Interaction Enjoyment The chatbot is perceived as enjoyable and engaging to operate with</p> <p>17. Personality Chatbot conveys a personality by providing greetings, self-introductory, empathy, information, etc.</p>

## Appendix 7. Chatbots And Tasks

### 1. AMTRAK—<https://www.amtrak.com/home>

**Task description:** 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.

### 2. TOSHIBA—<http://www.toshiba.co.uk/generic/yoko-home/>

**Task description:** 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 organise your video and audio libraries.

### 3. ATO—<https://www.ato.gov.au/>

**Task description:** 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.

### 4. INBENTA—<https://www.inbenta.com/en/>

**Task description:** 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.

### 5. 1-800-FLOWER ASSISTANT—<https://www.facebook.com/messages/t/1800FlowersAssistant>

**Task description:** 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.

### 6. HSBC UK—<https://www.hsbc.co.uk/>

**Task description:** 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.

### 7. ABSOLUT—<https://www.absolut.com/en/>

**Task description:** 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.

### 8. BOOKING.COM—<https://www.facebook.com/messages/t/131840030178250>

**Task description:** You are travelling to London from 5th July to 9th July with your family. You want to use [booking.com](https://www.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.

### 9. USICS—<https://www.uscis.gov/emma>

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

### 10. TOMMY HILFIGER—<https://www.messenger.com/t/tommyhilfiger>

**Task description:** You bought a bottle of 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.

## Appendix 8 Bot Usability Scale (15 Items)

The present version of the BUS was developed for further testing. The current version was tested with a five-point Likert scale from 1 ('Strongly Disagree') to 5 ('Strongly Agree')

Factor	Item
1 - Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable. 2. It was easy to find the chatbot.
2 - Perceived quality of chatbot functions	3. Communicating with the chatbot was clear. 4. I was immediately made aware of what information the chatbot can give me. 5. The interaction with the chatbot felt like an ongoing conversation. 6. The chatbot was able to keep track of context. <sup>14</sup> 7. The chatbot was able to make references to the website or service when appropriate. 8. The chatbot could handle situations in which the line of conversation was not clear. 9. The chatbot's responses were easy to understand.
3 - Perceived quality of conversation and information provided	10. I find that the chatbot understands what I want and helps me achieve my goal. 11. The chatbot gives me the appropriate amount of information. 12. The chatbot only gives me the information I need. 13. I feel like the chatbot's responses were accurate.
4 - Perceived privacy and security	14. I believe the chatbot informs me of any possible privacy issues.
5 - Time response	15. My waiting time for a response from the chatbot was short.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Amazon. (2018). *Voice Design Checklist*. Amazon. Retrieved 26<sup>th</sup> April from <https://developer.amazon.com/docs/alexa-design/checklists.html#voice-design-checklist>
- Applin SA, Fischer MD (2015) *New technologies and mixed-use convergence: How humans and algorithms are adapting to each other*. IEEE International Symposium on Technology and Society (ISTAS), Dublin
- Borsci S, Federici S, Bacci S, Gnaldi M, Bartolucci F (2015) Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction* 31(8):484–495. <https://doi.org/10.1080/10447318.2015.1064648>
- Brandtzaeg, P. B., & Følstad, A. (2017). *Why people use chatbots* International Conference on Internet Science
- Brandtzaeg PB, 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
- Businesswire. (2019). *The Global Chatbot Market is Forecast to Reach \$5.63 Billion by 2023 – Asia Pacific to Witness the Highest Growth - ResearchAndMarkets.com*. Businesswire. Retrieved 20th of August from <https://www.businesswire.com/news/home/20190314005364/en/Global-Chatbot-Market-Forecast-Reach-5.63-Billion>
- Caruana A (2002) Service loyalty. *European Journal of Marketing* 36(7/8):811–828. <https://doi.org/10.1108/03090560210430818>
- Charlton JI (2000) *Nothing about us without us: Disability oppression and empowerment*. Univ of California Press
- Chaves, A. P., & Gerosa, M. A. (2019). How should my chatbot interact? A survey on human-chatbot interaction design. *arXiv preprint arXiv:1904.02743*.
- Chopra, S., & Chivukula, S. (2017). *My phone assistant should know I am an Indian: influencing factors for adoption of assistive agents* Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, Vienna, Austria.

12. Chung H, Iorga M, Voas J, Lee S (2017) Alexa, can I trust you? *Computer* 50(9):100–104
13. Chung M, Ko E, Joung H, Kim SJ (2020) Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research* 117:587–595. <https://doi.org/10.1016/j.jbusres.2018.10.004>
14. Cohen, D., & Lane, I. (2016). *An oral exam for measuring a dialog system's capabilities* 13<sup>th</sup> AAAI Conference on Artificial Intelligence, Phoenix, US.
15. Coniam D (2014) The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk* 34(5):545–567
16. Conti G, Frühwirth-Schnatter S, Heckman JJ, Piatek R (2014) Bayesian exploratory factor analysis. *Journal of econometrics* 183(1):31–57
17. Dale R (2016) The return of the chatbots. *Natural Language Engineering* 22(5):811–817
18. De Souza CS (2005) *The semiotic engineering of human-computer interaction*. MIT press
19. De Souza CS, Leitão CF (2009) Semiotic engineering methods for scientific research in HCI. *Synthesis Lectures on Human-Centered Informatics* 2(1):1–122
20. Derrick DC, Meservy TO, Jenkins JL, Burgoon JK, Nunamaker JF Jr (2013) Detecting deceptive chat-based communication using typing behavior and message cues. *ACM Transactions on Management Information Systems (TMIS)* 4(2):9
21. DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.
22. Dillon A (2001) Beyond Usability: Process, Outcome and Affect in human computer interactions. *Canadian Journal of Information and Library Science* 26(4)
23. Federici, S., de Filippis, M. L., Mele, M. L., Borsci, S., Bracalenti, M., Gaudino, G., Cocco, A., Amendola, M., & Simonetti, E. (2020). Inside pandora's box: a systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. Disability and Rehabilitation: Assistive Technology, 1–6. <https://doi.org/10.1080/17483107.2020.1775313>,
24. Feine, J., Morana, S., & Gnewuch, U. (2019). *Measuring service encounter satisfaction with customer service chatbots using sentiment analysis* 14th International Conference on Wirtschaftsinformatik (WI2019), Siegen, Germany. <https://aisel.aisnet.org/wi2019/>
25. Finstad K (2010) The usability metric for user experience. *Interacting with Computers* 22(5):323–327
26. Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 4(2):e19. <https://doi.org/10.2196/mental.7785>
27. Følstad A, Brandtzaeg PB (2017) Chatbots and the new world of HCI. *Interactions* 24(4):38–42
28. Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2019). Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design International Conference on Internet Science, Cham.
29. Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? SIGCHI conference on Human Factors in Computing Systems, The Hague, The Netherlands.
30. Gaudio, P., & Kater, K. (2000). ALife-WebGuide: an intelligent user interface for Web site navigation Proceedings of the 5th international conference on Intelligent user interfaces, New Orleans, Louisiana, USA.
31. Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service.
32. Go E, Sundar SS (2019) Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97:304–316. <https://doi.org/10.1016/j.chb.2019.01.020>
33. Google. (2017). *Learn about conversation*. Google. Retrieved 13th of March from <https://designguidelines.withgoogle.com/conversation/conversation-design/learn-about-conversation.html#learn-about-conversation-the-cooperative-principle>
34. Harkous, H., Fawaz, K., Shin, K. G., & Aberer, K. (2016). Pribots: Conversational privacy with chatbots 12<sup>th</sup> Symposium on Usable Privacy and Security, Denver, Colorado, US.
35. Hayes AF, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1(1):77–89
36. Hertzum M, Andersen HH, Andersen V, Hansen CB (2002) Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers* 14(5):575–599
37. Hinrichs, H., & Le, N.-T. (2018). Which text-mining technique would detect most accurate user frustration in chats with conversational agents? 32<sup>nd</sup> International BCS Human Computer Interaction Conference, Belfast, United Kingdom.
38. Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI).
39. Hoofs H, van de Schoot R, Jansen NWH, Kant I (2018) Evaluating Model Fit in Bayesian Confirmatory Factor Analysis With Large Samples: Simulation Study Introducing the BRMSEA. *Educational and psychological measurement* 78(4):537–568. <https://doi.org/10.1177/0013164417709314>
40. Huang, C.-Y., & Ku, L.-W. (2018). EmotionPush: Emotion and Response Time Prediction Towards Human-Like Chatbots IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates.
41. IBM Conversational UX (2018). Talk meets technology - Conversation design guidelines. IBM. Retrieved 2nd May from <http://conversational-ux.mybluemix.net/design/conversational-ux/practices/>
42. Io, H. N., & Lee, C. B. (2017, 10-13 Dec. 2017). Chatbots and conversational agents: A bibliometric analysis. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),
43. ISO. (2018). ISO 9241-11:2018 Ergonomic requirements for office work with visual display terminals – Part 11: Guidance on usability. In. Brussels, BE: CEN.
44. Ives B, Olson MH, Baroudi JJ (1983) The measurement of user information satisfaction. *Communications of the ACM* 26(10): 785–793
45. Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots Designing Interactive Systems Conference, Hong Kong.
46. Kirakowski J, O'Donnell P, Yiu A (2009) Establishing the hallmarks of a convincing chatbot-human dialogue. In: Maurtua I (ed) Human-Computer Interaction. IntechOpen, pp 49–56. <https://doi.org/10.5772/7741>
47. Kuligowska K (2015) Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research* 2(2):1–16. <https://doi.org/10.18483/PCBR.22>
48. Lee S, Choi J (2017) Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103:95–105. <https://doi.org/10.1016/j.ijhcs.2017.02.005>
49. Lemon, O., Gruenstein, A., Battle, A., & Peters, S. (2002). Multi-tasking and collaborative activities in dialogue systems 3<sup>rd</sup> SIGDial Workshop on Discourse and Dialogue, Philadelphia, US.

50. Lewis, J. R. (2001). Psychometric properties of the mean opinion scale. *Proceedings of HCI International 2001: Usability Evaluation and Interface Design*, 149–153.
51. Lewis JR (2019) Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter? *Human Factors, Online first* 0018720819881312:001872081988131. <https://doi.org/10.1177/0018720819881312>
52. Lewis JR, Hardzinski ML (2015) Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology* 18(3):479–487. <https://doi.org/10.1007/s10772-015-9289-1>
53. Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: when there's no time for the SUS SIGCHI Conference on Human Factors in Computing Systems, Paris, France.
54. Lindgaard, G., & Dudek, C. (2002). User Satisfaction, Aesthetics and Usability: Beyond Reductionism IFIP 17<sup>th</sup> World Computer Congress - TC13 Stream on Usability: Gaining a Competitive Edge, Deventer, The Netherlands.
55. Mcknight DH, Carter M, Thatcher JB, Clay PF (2011) Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)* 2(2):12:12–12:25. <https://doi.org/10.1145/1985347.1985353>
56. McTear MF, Callejas Z, Griol D (2016) Speech Input and Output. In: McTear MF, Callejas Z, Griol D (eds) *The Conversational Interface Talking to Smart Devices*. Springer, pp 75–92. [https://doi.org/10.1007/978-3-319-32967-3\\_5](https://doi.org/10.1007/978-3-319-32967-3_5)
57. Meira, M., & Canuto, A. D. P. (2015). Evaluation of emotional agents' architectures: an approach based on quality metrics and the influence of emotions on users World congress on engineering (WCE 2015), London, UK.
58. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E (2016) Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine* 176(5):619–625
59. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine* 151(4):264–269
60. Morrissey, K., & Kirakowski, J. (2013). 'Realness' in Chatbots: *Establishing Quantifiable Criteria* International Conference on Human-Computer Interaction,
61. Munteanu, C., & Boldea, M. (2000). *MDWOZ: A Wizard of Oz Environment for Dialog Systems Development* International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece.
62. Paikari, E., & van der Hoek, A. (2018). *A framework for understanding chatbots and their future* The 11<sup>th</sup> International Workshop on Cooperative and Human Aspects of Software Engineering, Gothenburg, Sweden.
63. Pauletto S, Balentine B, Pidcock C, Jones K, Bottaci L, Aretoulaki M, Wells J, Mundy DP, Balentine J (2013) Exploring expressivity and emotion with artificial voice and speech technologies. *Logopedics Phoniatrics Vocology* 38(3):115–125
64. Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, 89–97
65. Pereira J, Diaz Ó (2019) What Matters for Chatbots? Analyzing Quality Measures for Facebook Messenger's 100 Most Popular Chatbots. In: Majchrzak TA, Mateos C, Poggi F, Grønli T-M (eds) *Towards Integrated Web, Mobile, and IoT Technology*. Springer International Publishing, pp 67–82
66. Piatek, R. (2017). BayesFM: Bayesian Inference for Factor Modeling. R package version 0.1, 2.
67. Polkosky, M. D. (2008). Machines as mediators: The challenge of technology for interpersonal communication theory and research. In *Mediated interpersonal communication* (pp. 48–71). Routledge.
68. Quarteroni S, Manandhar S (2009) Designing an interactive open-domain question answering system. *Natural Language Engineering* 15(1):73–95
69. Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv: 1704.04579.
70. Revelle W (2011) An overview of the psych package. *Department of Psychology Northwestern University*. Accessed on March 3(2012):1–25
71. Saenz, J., Burgess, W., Gustitis, E., Mena, A., & Sasangohar, F. (2017). The usability analysis of chatbot technologies for internal personnel communications Industrial and Systems Engineering Conference Pittsburgh, Pennsylvania, US.
72. Sankar, G. R., Greyling, J., Vogts, D., & Plessis, M. C. D. (2008). Models towards a hybrid conversational agent for contact centres Annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology, Wilderness, South Africa.
73. Sauro, J. (2017). *Measuring usability: From the SUS to the UMUX-LITE*. measuringU. Retrieved 3rd February from <https://measuringu.com/umux-lite/>
74. Staven, T. (2017). *What Makes a Good Bot or Not?* unit4. Retrieved 15th of April from <https://www.unit4.com/blog/2017/03/what-makes-a-good-bot-or-not>
75. Tay L, Jebb A (2017) Scale Creation. In: Rogelberg SG (ed) *The SAGE encyclopedia of industrial and organizational psychology*, 2nd edn. SAGE Publications, Inc.
76. Thorne C (2017) Chatbots for troubleshooting: A survey. *Language and Linguistics Compass* 11(10):e12253. <https://doi.org/10.1111/lnc3.12253>
77. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB (2019) Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry* 64(7):456–464. <https://doi.org/10.1177/0706743719828977>
78. Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2017). *Here's What I Can Do: Chatbots' Strategies to Convey Their Features to Users* The XVI Brazilian Symposium on Human Factors in Computing Systems, Joinville, Brazil.
79. Van Eeuwen, M. (2017). *Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers* University of Twente].
80. Verhagen T, van Nes J, Feldberg F, van Dolen W (2014) Virtual Customer Service Agents: Using Social Presence and Personalization to Shape Online Service Encounters. *Journal of Computer-Mediated Communication* 19(3):529–545. <https://doi.org/10.1111/jcc4.12066>
81. Vetter M (2002) Quality Aspects of Bots. In: Meyerhoff D, Laibarra B, van der Pouw Kraan R, Wallet A (eds) *Software Quality and Software Testing in Internet Times*. Springer, Berlin, pp 165–184. [https://doi.org/10.1007/978-3-642-56333-1\\_11](https://doi.org/10.1007/978-3-642-56333-1_11)



82. Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. arXiv preprint [cmp-lg/9704004](https://arxiv.org/abs/cmp-lg/9704004).
83. Weiss A, Tscheligi M (2012) Rethinking the Human–Agent Relationship: Which Social Cues Do Interactive Agents Really Need to Have? In: Hingston P (ed) *Believable Bots: Can Computers Play Like People?* Springer, Berlin, pp 1–28. [https://doi.org/10.1007/978-3-642-32323-2\\_1](https://doi.org/10.1007/978-3-642-32323-2_1)
84. Wilson HJ, Daugherty P, Bianzino N (2017) The jobs that artificial intelligence will create. *MIT Sloan Management Review* 58(4):14
85. Yang, F., Heeman, P. A., & Kun, A. (2008). *Switching to real-time tasks in multi-tasking dialogue* 22<sup>nd</sup> International Conference on Computational Linguistics, Manchester, UK.
86. Yang, X., Oikarinen, E.-L., & Saraniemi, S. (2020). Understanding Chatbot Service Encounters: Consumers' satisfactory and dissatisfactory experiences University of Oulu]. Oulu Business School.
87. Zamora, J. (2017). *I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations* 5<sup>th</sup> International Conference on Human Agent Interaction, Bielefeld, Germany.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.