Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-022-01690-0

ORIGINAL PAPER

A confirmatory factorial analysis of the Chatbot Usability Scale: a multilanguage validation

Simone Borsci[1,2] · Martin Schmettow[1] · Alessio Malizia[3,4] · Alan Chamberlain[5] · Frank van der Velde[1]

Received: 4 October 2021 / Accepted: 12 July 2022

Abstract

The Bot Usability Scale (BUS) is a standardised tool to assess and compare the satisfaction of users after interacting with chatbots to support the development of usable conversational systems. The English version of the 15-item BUS scale (BUS-15) was the result of an exploratory factorial analysis; a confirmatory factorial analysis tests the replicability of the initial model and further explores the properties of the scale aiming to optimise this tool seeking for the stability of the original model, the potential reduction of items, and testing multiple language versions of the scale. BUS-15 and the usability metrics for user experience (UMUX-LITE), used here for convergent validity purposes, were translated from English to Spanish, German, and Dutch. A total of 1292 questionnaires were completed in multiple languages; these were collected from 209 participants interacting with an overall pool of 26 chatbots. BUS-15 was acceptably reliable; however, a shorter and more reliable solution with 11 items (BUS-11) emerged from the data. The satisfaction ratings obtained with the translated version of BUS-11 were not significantly different from the original version in English, suggesting that the BUS-11 could be used in multiple languages. The results also suggested that the age of participants seems to affect the evaluation when using the scale, with older participants significantly rating the chatbots as less satisfactory, when compared to younger participants. In line with the expectations, based on reliability, BUS-11 positively correlates with UMUX-LITE scale. The new version of the scale (BUS-11) aims to facilitate the evaluation with chatbots, and its diffusion could help practitioners to compare the performances and benchmark chatbots during the product assessment stage. This tool could be a way to harmonise and enable comparability in the field of human and conversational agent interaction.

Keywords Chatbots · Conversational agents · Usability · Artificial Intelligence · Interaction

1 Introduction

Chatbots for customer relationship management (CRM) are intended as intelligent conversational applications that can assist users in decision making through text (or voice) input and output [30, 33]. CRM chatbots are usually developed and adapted by the service provider to enable 24/7 rapid exchanges with potential customers. In this sense, CRM chatbots can vary substantially in terms of appearance, behaviour, and capabilities, providing a different experience to end-users [10].

As indicated by the ISO 9241–210 [24], a central aspect of the user experience (UX) is the satisfaction of the end-users defined as "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".

Satisfaction is a complex measure of the end-user reaction to and their reasoning about systems relating to the efficiency and effectiveness, and accuracy and reliability of assessment modalities [3, 13, 20, 25, 29].

User satisfaction is generally assessed after interaction with a given system, by using reliable usability scales such as the System Usability Scale (SUS, [7] and its shorter proxies, the Usability Metric for User Experience (UMUX,

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

[1] Department of Learning, Data Analysis, and Technology, Cognition, Data and Education (CODE) Group, Faculty of Behavioural Management and Social Sciences, University of Twente, Enschede, The Netherlands

[2] Department of Surgery and Cancer, Faculty of Medicine, NIHR London IVD, Imperial College of London, London, UK

[3] Computer Science Department, University of Pisa, Pisa, Italy

[4] Molde University College, Molde, Norway

[5] School of Computer Science, University of Nottingham, Nottingham, UK

Published online: 04 August 2022

[18]) and UMUX LITE [28]. Instead of questions about user satisfaction and ease of use, that are barely comparable [2, 3], standardised scales aim to assess the users' perspective after interacting with products, usually on a score from 0 to 100, to provide comparable insights regarding the quality of tools, by investigating the participants' perception of, and reaction to key interactive aspects of such experiences. Such standardised subjective assessment, when coupled with objective measures of effectiveness and efficiency can provide relevant, replicable, and comparable information about the usability (ISO 9241–11). Moreover, when this is used in conjunction with data collected over time in the context of use, the expected value and acceptance in respect to the satisfaction measures can support user experience researchers in their efforts to model the overall experience of people developing a given product [4].

Nevertheless, unlike classic interactive systems based on graphical elements, chatbots rely on textual and conversational aspects to engage the end-users [38], co-constructing the interaction and the meaning of the conversation [12]. In this sense, chatbots in many respects create a new paradigm in human–computer interaction by placing the conversational exchange at the centre and the interaction between the user and the technology [19]. Therefore, the assessment of the chatbots' end-user satisfaction should also consider aspects that are not usually included in the classic satisfaction evaluation, e.g. the quality of the conversational exchange.

As reported by Borsci et al. [5] when reviewing the domain of chatbots, it is the case that little is known about how to evaluate the end-user's perception of quality when interacting with chatbots. There is a growing interest in understanding how to assess and improve the interaction with such systems [16, 22, 31]; however, to our knowledge, there are currently no standardised tools to assess then end-user's satisfaction with chatbots, except for the recently developed ChatBot Usability Scale (BUS-15) [5]. The BUS-15 scale was developed and tested using an exploratory factorial analysis. This scale was developed by proposing an initial model of 42 items. It was developed via a systematic literature review, and by interviewing and surveying designers and users of chatbots. The exploratory analysis of the initial model of 42 items (i.e. key aspects associated with the experience with chatbots) resulted in 15 items divided into 5 factors (Table 1) with an overall reliability of 0.87, with factor 1 being composed of two items and Cronbach's alpha equal to 0.87, factor 2 composed of seven items and reliability equal to 0.74, and factor 3 composed by three items with an alpha value of 0.86; factors 4 and 5 were composed of single items. Moreover, The BUS-15 factors strongly correlate with UMUX-LITE (between 0.61 and 0.87), suggesting that the BUS-15 is reliable when used to assess the end-user's overall satisfaction and it also adds new elements not considered by classic satisfaction scales.

The present work aims to perform a confirmatory factorial analysis on BUS-15, testing its psychometric properties and potential alternative factorial models. The confirmatory analysis is considered a necessary step [32, 40] to validate a new scale by statistically checking and optimising the factorial model that emerged in the exploratory analysis [3]. In addition, we also conduct an analysis under a *designometric* perspective.

Table 1 The original (English) version of BUS 15. Each item is assessed on a 5-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
	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

Designometrics has recently been introduced by Schmettow [36], noting that the purpose of a user experience self-report scale (such as BUS or UMUX-LITE) is to compare designs, whereas psychometrics is focused on people. While the statistical workflow is the same, the data and the interpretation differ. A psychometric analysis of reliability requires multiple persons to respond to multiple items, which is often referred to as the psychometric response matrix. In designometric assessment the data collection must also include a *sample of designs*, resulting in a three-dimensional response matrix, which can be reduced to a design-by-item matrix to fit standard psychometric tools (such as reliability scores, exploratory, and confirmatory factorial analysis). The interpretation of designometric analysis refers to designs, rather than to people. If a chatbot satisfaction scale has a good designometric reliability, this means it measures very precisely how well the chatbot can lead to a high degree of user satisfaction. In contrast, under a psychometric perspective, the same scale measures how easily individual users are satisfied by any chatbot. Obviously, these interpretations are not the same and the second could be considered less relevant in the context of interactional assessment; Schmettow [36] even went as far as calling this a *psychometric fallacy*, if designometric scales are validated only under a psychometric perspective. In respect of this, we take a stance which presents both perspectives on two grounds: firstly, psychometric evaluation of user experience scales is mainstream, and we aim for compatibility with existing research. Secondly, in the present case, the focus is on the structure of a multi-scale inventory (BUS-15). We predict that the partitioning of items into multiple scales is dominated by mental processes, hence would produce similar results under both perspectives.

Moreover, additional aims of this work are (i) to test the validity of three versions of the BUS, translated by native speakers, in Spanish, Dutch, and German, and (ii) to investigate the correlation (convergent validity) between the BUS scale and the UMUX-LITE that was identified in the previous study of Borsci et al. [5].

2 Methods

2.1 Participants

The study was approved by the ethical committee of the University of Twente; it was advertised by a specialised service and using social media aiming to target a pool of international potential users of different ages. The sampling strategy was, by convenience, the only inclusion criteria specified in the advertisement was that participants should be proficient in writing and reading English to take part in the study.

Table 2 Number of questionnaires (BUS-15) per each available language: English, German, Dutch, Spanish

Bot Usability Scale version	Number of completed BUS 15
English	356
German	400
Dutch	426
Spanish	110

Each participant evaluated the interaction experience with a minimum of five to a maximum of ten chatbots, resulting in a total of 1292 completed questionnaires in multiple languages (English, German, Dutch, and Spanish) as reported in Table 2.

A total of 259 people participated in the testing of the scale, of these only 80.7% (209) participants filled the survey correctly—i.e. 128 female, 131 male, age average: 37.78 min 18, max 83, 89% of the participants were European. Fifty-four percent of the sample was under 40 years old, while the remaining part of the sample was over 40 years old. The 20.3% of participants were excluded because they decided to stop the assessment for personal issues, or they had technical problems and were not willing to continue the evaluation. In some cases, less than ten chatbots were correctly displayed to participants for different reasons, e.g. issues in the availability of the chatbots, issues due to internet connections, etc. In such cases, we retained the answer of the participants only when the answer to at least 5 chatbots was collected correctly.

2.2 Materials

The study was designed to enable participants to interact with chatbots and answer the questionnaire by using a survey developed with Qualtrics software. At the beginning of the survey, participants were asked to declare their nationality and native language. If their native language was Dutch, German, or Spanish, the participants were assigned to answer the questionnaire in one of these languages. If participants were not native in one of these three languages, they were asked to fill out the questionnaire in English. Participants were informed (see instructions in Appendix A) that most of the chatbots were mainly in English, and when chatbots were available in the other languages (Dutch, German, or Spanish) these were also presented in the native language of the participants. Each participant was asked to assess ten chatbots extrapolated from the list of 26 (see the list in Appendix A) by performing a task of information retrieval, e.g. find specific information to inform their decision making (see an example in Appendix A). The language capabilities of the chatbots were considered in the randomised presentation of the chatbot by allowing participants to also experience some of the chatbots in their native language. Moreover, participants were asked to fill a demographic section reporting two individual characteristics: age and sex.

The chatbot systems were selected among available CRM chatbots retrievable online and associated with a service offered by a company to guide their users in the process of information retrieval.

For each chatbot, after the interaction, the participants were asked to fill in the BUS-15 [5] and the two items of the UMUX-LITE [28]. The UMUX-LITE was presented to the participants on a scale with 5-Likert points instead of the classic 7-point commonly considered a safe reduction [27, 35]. The UMUX-LITE items were presented in one of the four languages. The process of translation and back translation of the questionnaires was performed by native speakers.

2.3 Procedure

After the introduction to the study, participants were asked to fill the demographic section. Then they were asked to interact with each chatbot to achieve a specific information retrieval task.

As the online survey was designed to ask participants to perform tasks with chatbots, each participant, when possible, performed the test by sharing their screen with a member of the research team. This procedure was done to offer support to the participants during the interaction and to ensure control over the gathered data. Researchers were instructed to only answer questions regarding potential misunderstanding or incomprehension related to the survey, and to mainly monitor that participants interacted with the chatbots.

When it was not possible to connect during the session (about 10% of the participants), a post-interview was performed to ask confirmation to the participants that they interacted with the chatbots and to ask about their difficulties in performing the tasks.

After each interaction with a chatbot, a total of seventeen questions (15 items from the BUS and the 2 items of the UMUX-LITE) were presented in a fully randomised order.

2.4 Data analysis

Data analysis was performed in R. A linear regression model was used to inspect whether there was a significant difference between the rating of satisfaction obtained with the translated version of the scales. For this analysis the BUS and the UMUX- LITE were used as the dependent variables, while the different languages of translations were used as the independent variable with the English version as the intercept.

The "lavaan" and "ggplot" packages of R were used for the confirmatory factorial analysis (CFA) with weighted least squares means and variance adjusted estimation. Factor loading was considered acceptable when at least 0.6 and optimal at 0.7 and above [21]. Model fit was established by looking at multiple criteria including [8, 23, 41] the ratio between chi-square and the degrees of freedom below 3; the comparative fit index (CFI) aiming for a value of 0.90 or higher; the root mean squared error approximation (RMSEA) aiming for values less than 0.07; the standardised root mean square residual (SRMR) looking for a value below 0.08.

Cronbach alpha was calculated for the overall scale and per each factor of the BUS 15. To establish convergent reliability, a Kendall tau correlation analysis was performed between BUS and UMUX-LITE. Finally, a regression analysis was performed to assess the differences among the different chatbots in terms of satisfaction measured by the BUS and the effect of individual characteristics on the satisfaction rated by participants with this scale.

3 Results

3.1 Validity of the translated version of the scales

A sub-sample of 503 questionnaires regarding 5 chatbots were collected using all the four available languages of the two scales, i.e. BUS-15 and UMUX-LITE. The regression analysis suggested that there are no significant differences among the three translated versions of BUS-15 and the original version in English; however, the participants who used the German version, on average, tended to rate the satisfaction with chatbots slightly lower when compared to the participants who used the other versions (see Fig. 1).

A significant effect was identified for the UMUX-LITE (Fig. 2) suggesting that the satisfaction ratings obtained with the Dutch and the German version of the UMUX negatively affect the overall rating of the participants' satisfaction ($R2 = 0.036$, $F(3, 500) = 5.83$, $p < 0.001$). Specifically, compared to the participants who used the original scale in English, participants who rated their satisfaction with the Dutch version of the scale

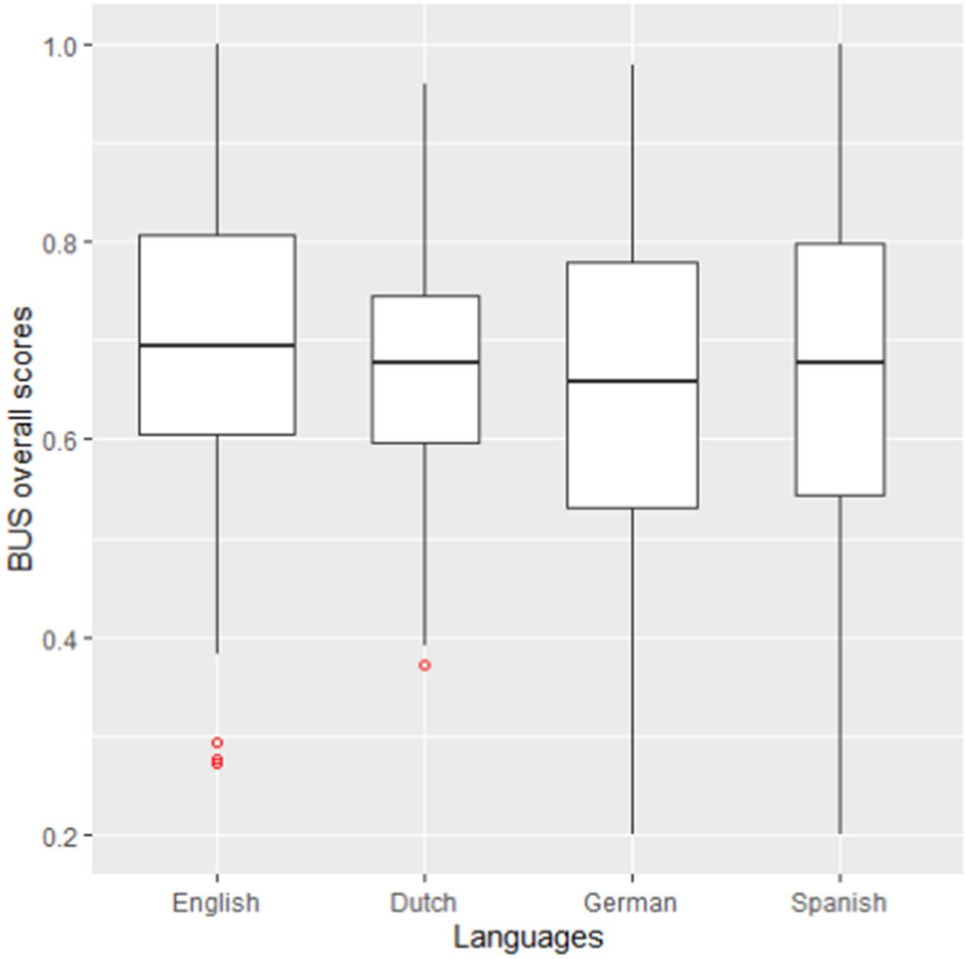

Fig. 1 Overall score of BUS 15 per each available language: English, Dutch, German, and Spanish

tended to report significantly lower satisfaction ratings ($b = -0.13$, $t(503) = -2.63$, $p = 0.05$). Similarly, participants who used the German version tended to rate their satisfaction lower than the other participants ($b = -0.11$, $t(503) = -3.860$, $p < 0.001$). This suggests that the Dutch and the German UMUX-LITE used in this study cannot be considered reliable for further analysis. Conversely, the original and Spanish versions (Appendix B) could be retained for further tests. The reliability of English UMUX-LITE ($\alpha = 0.89$) is higher than the one identified in previous studies (UMUX-LITE Cronbach's alpha between 0.82 and 0.83, [28] the Spanish version shows a comparable level of reliability ($\alpha = 0.83$).

3.2 The factor loading of the BUS-15

The results of the CFA is in line with the previous exploratory analysis [5] suggesting that the solution with five factors is acceptable with a CFI of 0.924 with loadings over the threshold of 0.6 (Table 3). The SRMR is equal to 0.039, and the RSMEA is equal to 0.065. The scale is strongly reliable with an overall Cronbach's alpha of 0.90.

Although the model appears robust (Fig. 3) the factors loading for items 4, 5, 7, and 8 are only acceptable, i.e. above but very close to 0.6. This might indicate that alternative models could be explored.

As reported in Table 4, two alternative factorial models were tested to optimise the scale. The first attempt was performed by removing the items with a barely acceptable factor loading, i.e. items 4, 5, 7, and 8. This resulted in a solution with 11 items (BUS-11) and five factors. The second alternative model was tested by also removing the factors with single items providing a solution with 9 items (BUS-9). This second model was tested because usually it is not considered an optimal solution to retain factors with less than three items [9, 15].

The BUS-9 is a short and very reliable solution with a Cronbach's alpha of 0.89; nevertheless, it is missing two aspects that emerged as relevant in interviews and focus groups in the original study [5], namely perceived

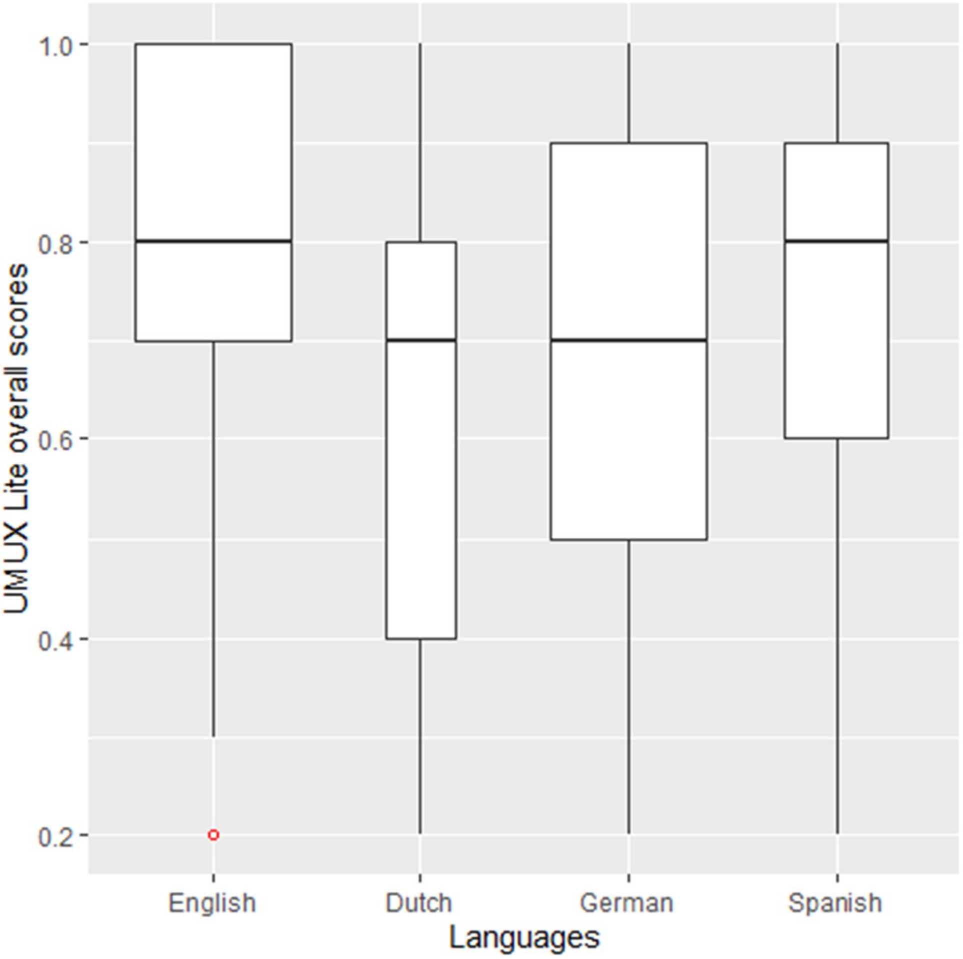

Fig. 2 Overall score of UMUX-LITE per each available language: English, Dutch, German, and Spanish

Table 3 Factor loading of BUS 15 with the original solution at 5 factors proposed by [5]

BUS 15 items	Factors				
	F1	F2	F3	F4	F5
Item 1	0.962				
Item 2	0.87				
Item 3		0.844			
Item 4		0.617			
Item 5		0.619			
Item 6		0.797			
Item 7		0.64			
Item 8		0.655			
Item 9		0.705			
Item 10			0.885		
Item 11			0.858		
Item 12			0.805		
Item 13			0.818		
Item 14				1	
Item 15					1
Reliability	0.90	0.87	0.91	-	-

privacy and security (factor 4) and time response (factor 5).

BUS-11 has a better fit than BUS 15, and it maintains the structure of original solutions while reducing the scale of four items (see Fig. 4). The overall reliability of the 11 items solution is relatively high ($\alpha = 0.89$) but the RSMEA on average is slightly over the expected threshold of 0.07.

The difference in terms of items between the two alternative models (BUS-11 and BUS-9) is minimal with the same overall reliability, but the BUS-11 is a more complete solution as this provides insights on specific and relevant characteristics of chatbots. Hence, the 11-item model (BUS-11) seems preferable to the shorter one (BUS-9).

The result of the CFA using the designometric matrix shows that while the model was stable in terms of factor loading for the first item of the scale, a problem due to negative variance, and the model fit was inferior compared to the psychometric model (chi-square/$df = 2.7$; CIF = 0.86, RSMEA = 0.27; SRMR = 0.06).

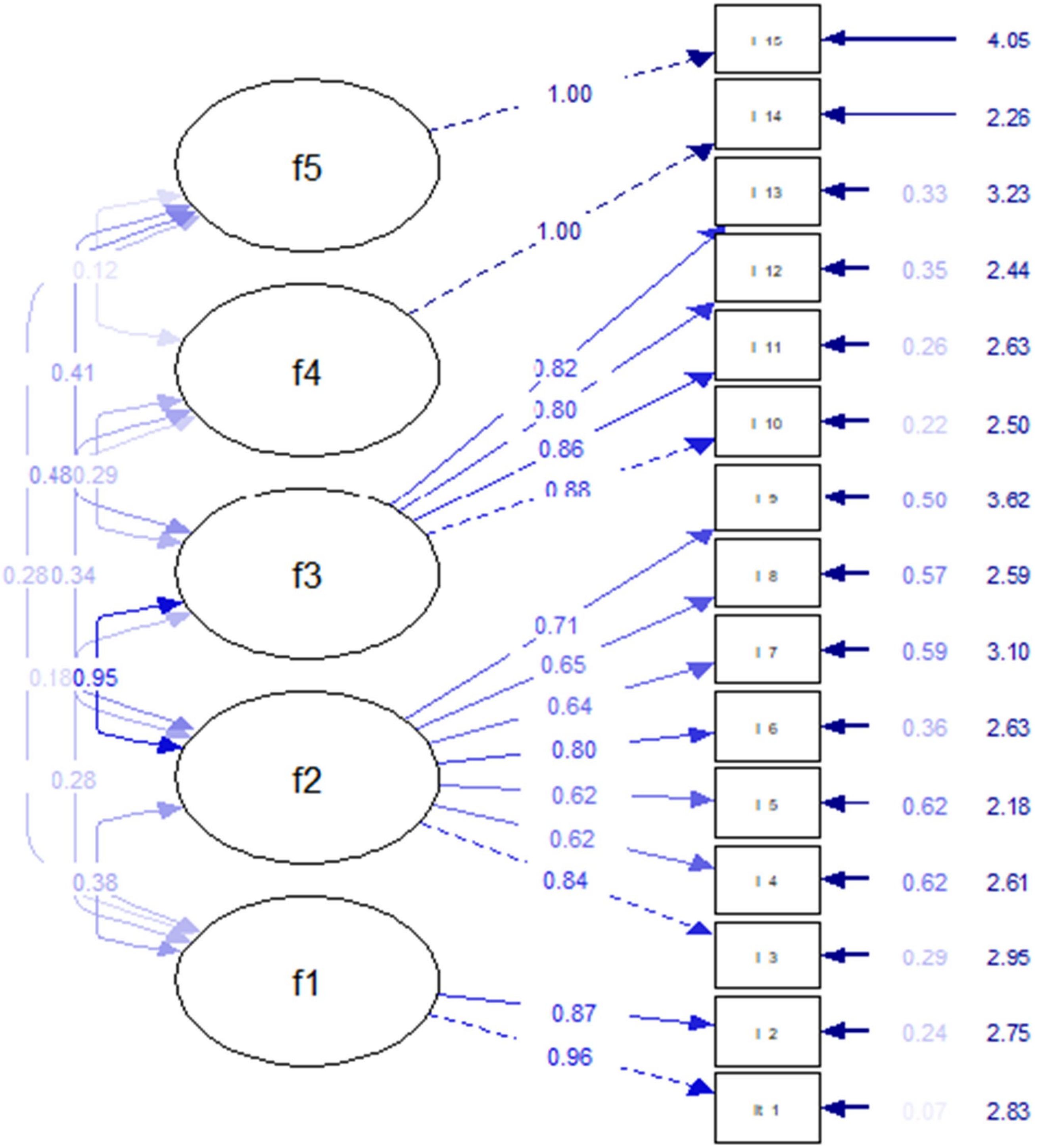

Fig. 3 Graphic representation of the factorial model of the BUS 15

3.3 Correlation between UMUX-LITE and BUS11 and effects of individual characteristics on BUS 11

The Kendall tau correlation analysis suggests that there is a positive significant relationship (rt = 0.68, $p < 0.001$) between the BUS-11 and the UMUX-LITE (English and Spanish versions) as shown in Fig. 5.

A linear model regression was performed to observe the difference in the satisfaction rating of the participant using BUS 11 after the interaction with the chatbots. Using a random chatbot as an intercept (here reported as C1) displayed

Table 4 Comparative analysis of the BUS-15 and two alternative models: BUS-11 and BUS-9

Fit indexes	Tested models		
	BUS- 15	BUS-11	BUS-9
Chi square/*df*	2.29	1.92	1.53
CIF	0.924	0.945	0.970
RSMEA	0.065 [CI 0.06–0.071]	0.071 [CI 0.063–0.079]	0.061 [CI 0.052–0.071]
SRMR	0.039	0.031	0.025

significant differences, except for chatbots 13, 18, and 19 (see Appendix C). Figure 6 shows the difference in the assessment of satisfaction among the chatbots.

The gender declared by participants does not affect the overall scores reported with BUS 11; however, the age of the participants has a moderate but significant effect ($R^2 = 0.008$, F (1, 1285) = 12.17, $p < 0.001$), suggesting that older participants were more conservative in their satisfaction rating towards chatbots compared to younger participants (b = –0.0009, t(1290) = − 3.48, $p < 0.001$).

4 Discussion

The 15-item model of the BUS, previously identified by Borsci et al. [5], is reliable but could be further optimised. We identified two solutions. The 9-item solution (BUS-9)

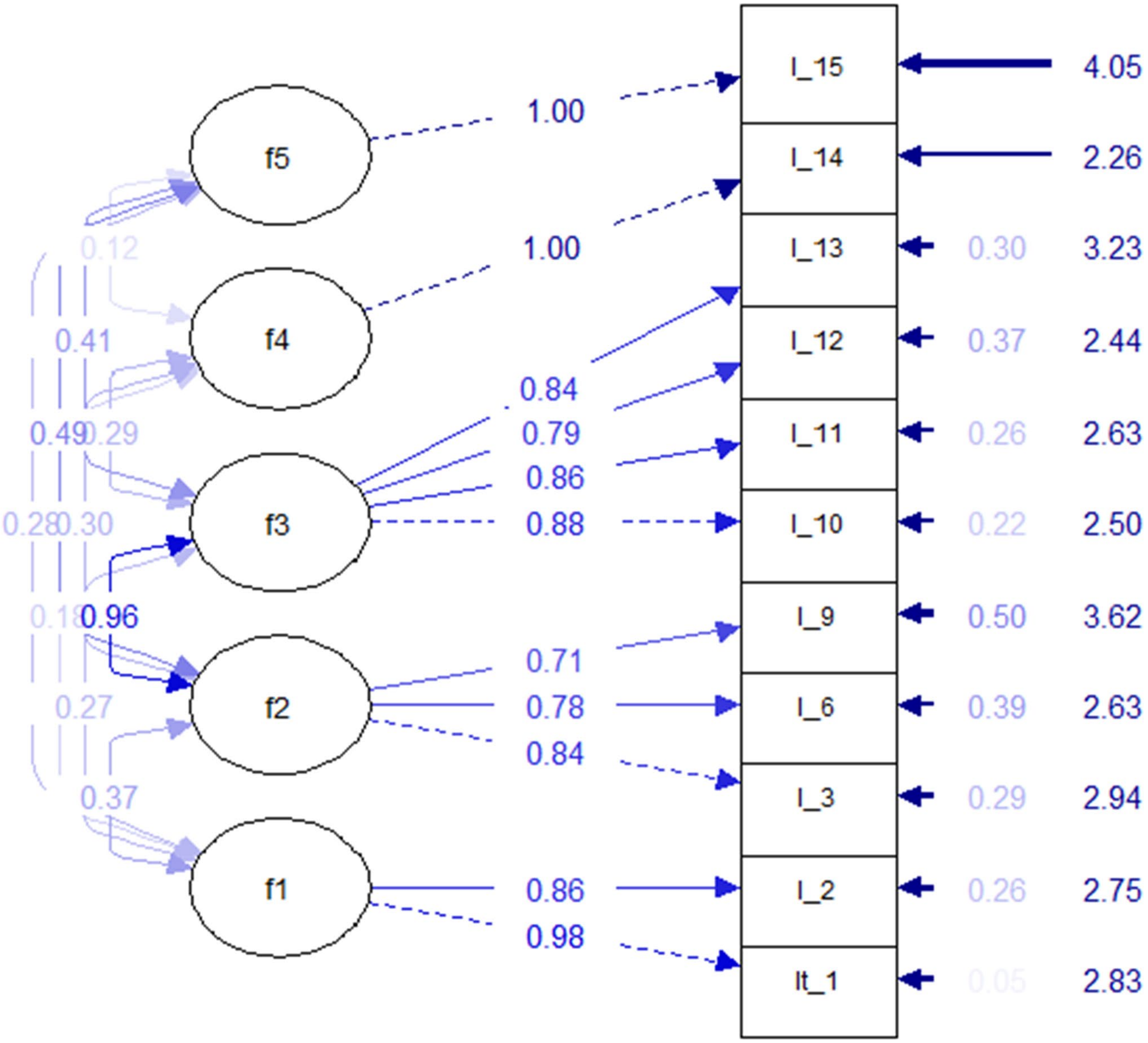

Fig. 4 Graphical representation of the factorial model of the BUS-11

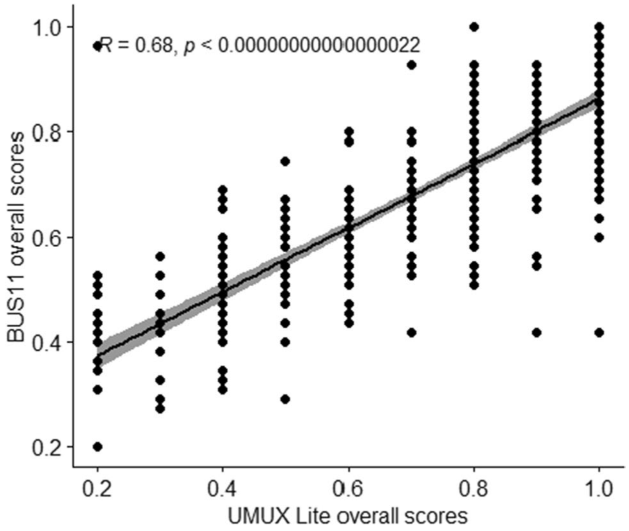

Fig. 5 Graphical representation of the correlation between the overall scores of the BUS-11 and the UMUX-LITE

offers a very reliable, short, and solid solution, but in such a case designers will lose relevant aspects to inform their decision making. Therefore, we recommend using BUS-11 also to collect data about key aspects such as privacy, security, and time to respond. However, BUS-9 could be used with mock-ups or early-stage prototypes when chatbots are not yet fully functional and specific aspects of the systems are still hard to judge for participants.

When comparing the 11 items modelled from a psychometric approach, using the 1259 answers of the participants, with the designometric perspective composed of 26 chatbots, BUS 11 appears to be stable in terms of factor and item organisation. Nevertheless, the designometric perspective resulted in an inferior fit with the factorial model as the CFI and SRMR indexes are acceptable while the RSMEA is particularly high, and for one item there is a problem of variance. These are common issues with small cohorts (W. R. [14, 26] and should be compensated by adding more chatbots to the database to completely model the ability of BUS 11 to differentiate critical design aspects of chatbots. However, looking at the psychometric and designometric analysis, the BUS-11 can be considered overall reliable. Moreover, the tool correlates with a classic satisfaction questionnaire (UMUX-LITE) by adding relevant elements regarding the specificity of the chatbots' functions, and it is now available in four languages: English, Dutch, German, and Spanish (Table 5).

The age of the users seems to affect the BUS-11 satisfaction rate of the participants slightly; this result is entirely in line with the results of a recent qualitative study which suggested that the age of the end-users is a relevant factor to assess the trustworthiness and interaction with chatbots [39]. Although the effect is minimal, it should be further explored in future studies.

The BUS-11 is a flexible scale used to assess user satisfaction of CRM chatbots; the current version of the tool has yet to be tested outside the domain of CRM and it could be considered providing a solid basis to investigate and support the design with other types of chatbots (e.g. general or domain-specific

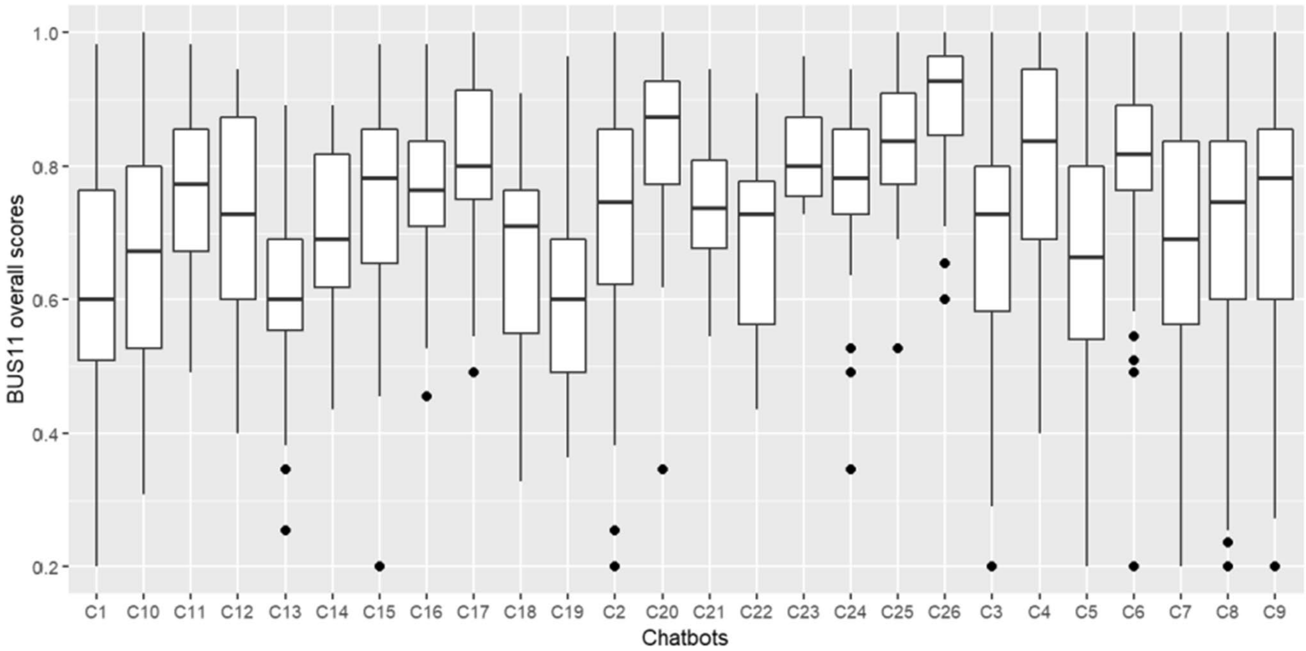

Fig. 6 Overall scores of BUS-11 per chatbots. Chatbots are numbered in an order different from the list presented in Appendix A as we did not ask permission or inform the service providers about the usage of these chatbots

Table 5 The BUS-11 versions are reported in English, Dutch, German, and Spanish. Each item is associated with a 5-point Likert scale from strongly agree (1) to strongly disagree (5)

Factors	English items	Dutch items	German items	Spanish items
1—Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable	1. De chatbot functie was makkelijk te ontdekken	1. Die Chatbot-Funktion war leicht zu erkennen	1. Pude reconocer la función del chatbot fácilmente
	2. It was easy to find the chatbot	2. Het was makkelijk om de chatbot te vinden	2. Es war einfach, den Chatbot zu finden	2. Fue fácil encontrar/localizar el chatbot
2—Perceived quality of chatbot functions	3. Communicating with the chatbot was clear	3. De communicatie met de chatbot was duidelijk	3. Die Kommunikation mit dem Chatbot war eindeutig	3. La comunicación con el chatbot fue clara
	4. The chatbot was able to keep track of context	4. De chatbot hield de context in het oog	4. Der Chatbot war in der Lage, den Kontext zu verfolgen	4. El chatbot pudo hacer el seguimiento del contexto de la conversación
	5. The chatbot's responses were easy to understand	5. De antwoorden van de chatbot waren gemakkelijk te begrijpen	5. Die Antworten des Chatbots waren einfach zu verstehen	5. Las respuestas del chatbot fueron fáciles de entender
3—Perceived quality of conversation and information provided	6. I find that the chatbot understands what I want and helps me achieve my goal	6. Ik denk dat de chatbot begrijpt wat ik wil en helpt me mijn doel te bereiken	6. Ich finde, dass der Chatbot versteht, was ich will und mir hilft, mein Ziel zu erreichen	6. Encuentro que el chatbot comprende lo que quiero y me ayuda a lograr mi objetivo
	7. The chatbot gives me the appropriate amount of information	7. De chatbot gaf me de juiste hoeveelheid informatie	7. Der Chatbot gibt mir die angemessene Menge an Informationen	7. El chatbot me da la cantidad adecuada de información
	8. The chatbot only gives me the information I need	8. De chatbot gaf me alleen de informatie die ik nodig had	8. Der Chatbot gibt mir nur die Informationen, die ich brauche	8. El chatbot solo me da la información que necesito
	9. I feel like the chatbot's responses were accurate	9. Ik had het gevoel dat de antwoorden van de chatbot klopten	9. Ich habe das Gefühl, dass die Antworten des Chatbots korrekt waren	9. Siento que las respuestas del chatbot fueron precisas
4—Perceived privacy and security	10. I believe the chatbot informs me of any possible privacy issues	10. Ik denk dat de chatbot me inlicht over mogelijke privacy problemen	10. Ich vertraue darauf, dass der Chatbot mich über mögliche Datenschutzprobleme informiert	10. Creo que el chatbot me informa sobre posibles problemas de privacidad
5—Time response	11. My waiting time for a response from the chatbot was short	11. Ik hoefde kort te wachten op een antwoord van de chatbot	11. Meine Wartezeit auf eine Antwort des Chatbots war kurz	11. Mi tiempo de espera para recibir una respuesta del chatbot fue breve

conversational agents) such as, for instance, tools for daily interaction, to support rehabilitation and adherence to medical treatment, etc. In such cases, we recommend using the BUS together with other reliable scales. Future studies will investigate the application of the BUS in other domains.

The present study has some limitations. First, by focusing on investigating the factorial structure of the BUS, participants were asked to interact with several chatbots; therefore, we minimised the questions about the individual characteristics. Future studies are going to explore which characteristics are affecting the satisfaction with chatbots measured with BUS 11 considering, for instance, education, expertise with chatbots, etc. [6]. Second, participants who had different native languages interacted with chatbots in English, despite this not affecting the goal of the study (i.e. testing the validity of the scale), this could have affected the interaction with the chatbots, and future studies should adopt a more naturalistic approach to testing chatbots available in the native language of the participants aiming to assess the perceived quality of the tools. Third, UMUX-LITE translation for the Dutch and the German versions was not optimal, as results suggest a significant difference with the English version. However, we proposed and preliminarily validated the Spanish version of UMUX-LITE that, to our knowledge, was not yet available in the literature. Future studies could use this version of the UMUX-LITE for further validation purposes. Finally, we used a small population of chatbots to build a three-dimensional perspective on the reliability of the BUS looking at the scale from the participants' (psychometric) and the chatbots' (designometric) perspectives. Despite the results are suggesting the overall construct of the BUS-11 is holding up when tested from the designometric, a perspective future study is under preparation where we are collecting data on additional chatbots to increase the sample of our "design" population.

5 Conclusion

The interest of practitioners regarding the usage of chatbots to support end-users of services is growing even in susceptible domains, including, for instance, e-government [17, 34] and health and rehabilitation [1, 16, 37]. As suggested by De Filippis et al. [11] and Federici et al. [16], there is a need for specific and calibrated tools to assess the quality of interaction of chatbots to support the designer during the development and the assessment of such systems. Borsci et al. [5] suggested that the quality of interaction with chatbots can only be ensured by defining reliable assessment criteria to ensure comparability and support a satisfactory interaction between people and these new types of intelligent technology.

The BUS-11 is a tool that can facilitate the evaluation of interaction with chatbots, and its diffusion could enable practitioners to compare the performances and benchmark their conversational systems during the formative and summative phase of product assessment. Concurrently, as Borsci et al. [5] proposed, designers could rely on a specific heuristic list, the BOT-check, to support their design thinking during the development phase of chatbots.

The new interactional paradigm shift created by chatbots is also opening a range of new research and design opportunities in the field of HCI [19], and the diffusion and usage of the BUS-11 could be a way to harmonise methods and ensure comparability of results.

Appendix A. Participant instructions, tasks, and list of chatbots

Instructions

In the next section you will be asked to interact with 10 different chatbots and performing a specific task (for example, find public transport subscription offers) and then to answer two questionnaires on the satisfaction of interacting with the chatbot.

In particular:

1. We will provide you with a link to a website which, once clicked, will open in a new browser page;
2. You will need to try to interact with the site chatbot to perform the required task. The tasks they have to perform are a way to collect information regarding the chatbot's functioning and the level of satisfaction of the interaction;
3. You will have to go back to the page of this questionnaire and fill in the two evaluation scales that you will find.

Please note that you are going to be assigned randomly to interact with a set of chatbots, it could be that you will interact only with English chatbots, or with a mix group of chatbots that can also talk also in your native language.

[Example of a task]

Remember that the purpose is to evaluate the satisfaction of interacting with the chatbot. If you succeed or not in completing the task below, when you think you have acquired enough information to be able to evaluate the interaction quality of the chatbot you can proceed to fill in the questionnaire.

Task:

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

Go to the website where you can find the chatbot of Amtrak: https://www.amtrak.com

Return to this page (by clicking on the relevant browser tab) when you believe that you have collected enough information to evaluate the chatbot.

	Chatbot link	Language capabilities
1	https://www.chatbot.com	English
2	https://www.utwente.nl/en/education/master/chat/?autostart=true	English
3	https://www.amtrak.com/home	English
4	https://www.lufthansa.com/digitalassistant/webchat.html	English
5	https://www.emiratesholidays.com	English
6	https://www.hdfcbank.com/personal/ways-to-bank	English
7	https://www.inbenta.com/	English and Spanish
8	https://www.benefitcosmetics.com	English
9	https://www.voegol.com	English, Spanish
10	https://www.absolut.com	English
11	https://www.amc.nl/	Dutch
12	https://www.bol.com/nl/nl/klantenservice/stel-je-vraag.html	Dutch
13	https://www.kpn.com/	Dutch
14	https://www.oxxio.nl/klantenservice	English, Dutch, Spanish, German
15	https://www.vattenfall.nl/	English, Dutch
16	https://www.asr.nl/	Dutch
17	https://www.ato.gov.au/	English
18	https://www.hsbc.co.uk/	English
19	https://www.uscis.gov/	English, Spanish
20	https://seattleballooning.com/	English
21	https://www.elster.de/eportal/start	German
22	https://www.congstar.de/	German
23	https://www.otto.de/	German
24	https://www.bahn.de/	German
25	https://wien.bot/	English, German
26	https://www.stadtwerke-troisdorf.de/	German

Appendix B. Spanish version of UMUX-LITE

1. Las capacidades de este sistema cumplen con mis requisitos
2. El sistema es fácil de usar

Appendix C

Chatbots	*B*	SE	*t*	*p*	CI_lower	CI_upper
C1 (intercept)	0.60	0.02	28.49	0.01**	0.56	0.64
C2	0.13	0.03	4.60	0.01**	0.07	0.18
C3	0.09	0.03	3.38	0.01**	0.04	0.14
C4	0.20	0.03	7.37	0.01**	0.15	0.26
C5	0.07	0.03	2.38	0.02*	0.01	0.12
C6	0.19	0.03	6.48	0.01**	0.14	0.25
C7	0.09	0.03	3.37	0.01**	0.04	0.14
C8	0.09	0.03	3.14	0.01**	0.04	0.15
C9	0.10	0.03	3.34	0.01**	0.04	0.16
C10	0.06	0.03	2.24	0.03*	0.01	0.11
C11	0.17	0.03	4.80	0.01**	0.10	0.23
C12	0.12	0.03	3.29	0.01**	0.05	0.18
C13	0.01	0.04	0.40	0.69	-0.06	0.08
C14	0.09	0.03	2.71	0.01**	0.03	0.16
C15	0.15	0.03	5.05	0.01**	0.09	0.21
C16	0.15	0.03	4.41	0.01**	0.09	0.22
C17	0.20	0.03	5.69	0.01**	0.13	0.27
C18	0.05	0.03	1.69	0.09*	-0.01	0.11
C19	0.00	0.03	-0.02	0.99	-0.06	0.06
C20	0.23	0.04	6.20	0.01**	0.16	0.30
C21	0.14	0.04	3.60	0.01**	0.07	0.22
C22	0.08	0.04	2.11	0.03*	0.01	0.16
C23	0.22	0.04	5.54	0.01**	0.14	0.30
C24	0.16	0.04	4.01	0.01**	0.08	0.24
C25	0.23	0.04	5.72	0.01**	0.15	0.30
C26	0.28	0.04	7.16	0.01**	0.21	0.36

Acknowledgements We would like to thank the students Imke Silderhuis, Mirre Van Bos, Jasmin Sophie Bigga, Niklas Pollmann, and Sabin Manuela Kerwien Lopez, for their help in gathering the data. 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 [grant number EP/ V00784X/1] UKRI Trustworthy Autonomous Systems Hub and [grant number EP/S035362/1] PETRAS 2.

References

1. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M (2019) An overview of the features of chatbots in mental health: a scoping review. Int J Med Informatics 132:103978
2. Borsci S, Buckle P, Walne S (2020) Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? Appl Ergon 84:103007. https://doi.org/10.1016/j.apergo.2019.103007
3. 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. Int J Human-Comput Inter 31(8):484–495. https://doi.org/10.1080/10447318.2015.1064648
4. Borsci S, Federici S, Malizia A, De Filippis ML (2019) Shaking the usability tree: why usability is not a dead end, and a constructive way forward. Behav Inform Technol 38(5):519–532. https://doi.org/10.1080/0144929X.2018.1541255
5. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, Chamberlain A (2021) The Chatbot Usability Scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput. https://doi.org/10.1007/s00779-021-01582-9
6. Brandtzaeg PB, Følstad A (2017) Why People Use Chatbots. In: Kompatsiaris I, Cave J, Satsiou A, Carle G, Passani A, Kontopoulos E, Diplaris S, McMillan D (eds) International conference on internet science. Springer International Publishing, pp 377–392
7. Brooke J (1996) SUS-A quick and dirty usability scale. Usabil Evaluat Indust 189(194):4–7
8. Cole DA (1987) Utility of confirmatory factor analysis in test validation research. J Consult Clin Psychol 55(4):584
9. Costello AB, Osborne J (2005) Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 10(1):7
10. Dale R (2016) The return of the chatbots. Nat Lang Eng 22(5):811–817
11. De Filippis ML, Federici S, Mele ML, Borsci S, Bracalenti M, Gaudino G, ..., Simonetti E (2020) Preliminary results of a systematic review: quality assessment of conversational agents (chatbots) for people with disabilities or special needs. Paper presented at the International Conference on Computers Helping People with Special Needs
12. Dev J, Camp LJ (2020) User engagement with chatbots: a discursive psychology approach. Paper presented at the Proceedings of the 2nd Conference on Conversational User Interfaces
13. Dillon A (2001) Beyond usability: process, outcome and affect in human computer interactions. Can J Inform Library Sci 26(4)
14. Dillon WR, Kumar A, Mulani N (1987) Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. Psychol Bull 101(1):126
15. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 4(3):272
16. Federici S, de Filippis ML, Mele ML, Borsci S, Bracalenti M, Gaudino G, ..., 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. Disabil Rehabil Assist Technol 15(7):832-837. https://doi.org/10.1080/17483107.2020.1775313
17. Federici S, Mele ML, Bracalenti M, De Filippis ML, Lanzilotti R, Desolda G, ... Simonetti E (2021) A Chatbot Solution for eGLU-Box Pro: The Usability Evaluation Platform for Italian Public Administrations. Paper presented at the Human-Computer Interaction. Theory, Methods and Tools, Cham
18. Finstad K (2010) The usability metric for user experience. Interact Comput 22(5):323–327
19 Følstad A, Brandtzæg PB (2017) Chatbots and the new world of HCI. Interactions 24(4):38–42
20. Frøkjær E, Hertzum M, Hornbæk K (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? Paper presented at the SIGCHI conference on Human Factors in Computing Systems, The Hague, The Netherlands
21. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2010) Multivariate data analysis, 7th edn. Prentice Hall, Upper Saddle River, New Jersey
22. Haugeland IKF, Følstad A, Taylor C, Alexander C (2022) Understanding the user experience of customer service chatbots: an experimental study of chatbot interaction design. Int J Hum Comput Stud 161:102788. https://doi.org/10.1016/j.ijhcs.2022.102788
23 Hu LT, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling 6(1):1–55
24. ISO (2019) ISO 9241–210 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. In. Brussels, BE: CEN
25. Ives B, Olson MH, Baroudi JJ (1983) The measurement of user information satisfaction. Commun ACM 26(10):785–793
26. Kenny DA, Kaniskan B, McCoach DB (2015) The performance of RMSEA in models with small degrees of freedom. Sociol Methods Res 44(3):486–507
27. Lewis JR (2019) Measuring user experience with 3, 5, 7, or 11 points: does it matter? Hum Factors 63:0018720819881312. https://doi.org/10.1177/0018720819881312
28. Lewis JR, Utesch BS, Maher DE (2013) UMUX-LITE: when there's no time for the SUS. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Paris, France
29. Lindgaard G, Dudek C (2002) User satisfaction, aesthetics and usability: beyond reductionism. Paper presented at the IFIP 17th World Computer Congress - TC13 Stream on Usability: Gaining a Competitive Edge, Deventer, The Netherlands
30. 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, Switzerland, pp 75–92
31. Nordheim CB, Følstad A, Bjørkli CA (2019) An initial model of trust in chatbots for customer service—findings from a questionnaire study. Interact Comput 31(3):317–335
32. Osborne JW, Fitzpatrick DC (2012) Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. Pract Assess Res Eval 17(1):15
33. Paikari E, van der Hoek A (2018) A framework for understanding chatbots and their future. Paper presented at the The 11th International Workshop on Cooperative and Human Aspects of Software Engineering, Gothenburg, Sweden
34. Portela M (2021) Interfacing participation in citizen science projects with conversational agents. Human Comput 8(2):33–53
35. Sauro J (2017) Measuring Usability: From the SUS to the UMUX-Lite. measuringu.com. Retrieved from https://measuringu.com/umux-lite/
36. Schmettow M (2021) New statistics for design researchers. Springer International Publishing, Cham
37. Su Z, Schneider JA, Young SD (2021) The role of conversational agents for substance use disorder in social distancing contexts. Subst

Use Misuse 56(11):1732–1735. https://doi.org/10.1080/10826084.2021.1949609

38. Valério FAM, Guimarães TG, Prates RO, Candello H (2018) Chatbots explain themselves: designers' strategies for conveying chatbot features to users. J Interact Syst 9(3)
39. van der Goot MJ, Pilgrim T (2020) Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. Paper presented at the Chatbot Research and Design, Cham
40. Van Prooijen J-W, Van Der Kloot WA (2001) Confirmatory analysis of exploratively obtained factor structures. Educ Psychol Measur 61(5):777–792
41. Wheaton B, Muthen B, Alwin DF, Summers GF (1977) Assessing reliability and stability in panel models. Sociol Methodol 8:84–136

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