



Ciao AI: the Italian adaptation and validation of the Chatbot Usability Scale

Simone Borsci^{1,2} · Elisa Prati³ · Alessio Malizia^{4,5} · Martin Schmettow¹ · Alan Chamberlain⁶ · Stefano Federici⁷

Received: 3 October 2022 / Accepted: 20 May 2023
© The Author(s) 2023

Abstract

Chatbot-based tools are becoming pervasive in multiple domains from commercial websites to rehabilitation applications. Only recently, an eleven-item satisfaction inventory was developed (the ChatBot Usability Scale, BUS-11) to help designers in the assessment process of their systems. The BUS-11 has been validated in multiple contexts and languages, i.e., English, German, Dutch, and Spanish. This scale forms a solid platform enabling designers to rapidly assess chatbots both during and after the design process. The present work aims to adapt and validate the BUS-11 inventory in Italian. A total of 1360 questionnaires were collected which related to a total of 10 Italian chatbot-based systems using the BUS-11 inventory and also using the lite version of the Usability Metrics for User eXperience for convergent validity purposes. The Italian version of the BUS-11 was adapted in terms of the wording of one item, and a Multi-Group Confirmatory Factorial Analysis was performed to establish the factorial structure of the scale and compare the effects of the wording adaptation. Results indicate that the adapted Italian version of the scale matches the expected factorial structure of the original scale. The Italian BUS-11 is highly reliable (Cronbach alpha: 0.921), and it correlates to other measures of satisfaction (e.g., UMUX-Lite, $\tau_b = 0.67$; $p < .001$) by also offering specific insights regarding the chatbots' characteristics. The Italian BUS-11 can be confidently used by chatbot designers to assess the satisfaction of their users during formative or summative tests.

Keywords Chatbots · Conversational agents · Usability · Satisfaction · User experience · Artificial intelligence · Language · Text

1 Introduction

The interest in chatbots as tools to enable mediated communication with users and ensure engagement during the interaction with systems is growing both in the scientific and business communities (Eisingerich et al., 2019; Rutz et al., 2019). The North America and European chatbot market was valued at USD 1826.3 million in 2020, and it is expected to constantly develop and grow in the next 10 years. Among these European countries, Italy is witnessing an increased adoption of chatbots for a range of different services [25], and, as suggested by the Digital Innovation Observatories of the Polytechnic of Milan, the market for artificial intelligence in Italy is in constant evolution [13] with a recent increase of the use of chatbots', which equates to an increased market size of 34% in Italy [27].

Experts [18, 23, 29] have recognized chatbots as systems of communication that are able to provide a beneficial service to end-users, allowing them to receive support and

✉ Simone Borsci
s.borsci@utwente.nl

¹ Faculty of Behavioural Management and Social Sciences, Department of Learning, Data Analysis, and Technology, Group of Cognition, Data and Education, University of Twente, Enschede, Netherlands

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

³ Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy

⁴ Computer Science Department, University of Pisa, Pisa, Italy

⁵ Molde University College, Molde, Norway

⁶ School of Computer Science, University of Nottingham, Nottingham, UK

⁷ Department of Philosophy, Social and Human Sciences and Education, University of Perugia, Perugia, Italy

information 24/7 and enabling companies to collect opinions from their consumers for marketing purposes. If we examine the Scopus database, over the last 10 years (in the English language), using the keyword “chatbots” and searching only within the title and abstracts of conferences and peer-reviewed journals (see Fig. 1), more than 2.847 journals and conference articles were published from 2012, with a peak in production of items from 2017 till today (about 2.764 articles). It is also not surprising to discover that without imposing a time limitation, the production of scientific articles about chatbots started in 2000 with a total of 53 until 2012.

As suggested by Adamopoulou and Moussiades [1], the recent focus of interest on such tools might be due to the advancement in the artificial intelligence (A.I.) domain and the ability to integrate chatbots into social media platforms. Moreover, this focus might also be due to the fact that these tools are considered flexible systems that can deliver more value to people, when compared to classic point-and-click systems, across a wider range of domains, e.g., education, rehabilitation, and consumer service [15, 17].

The value delivered by chatbots to their users, however, is not only associated with the possibility of interacting in a conversational modality with these digital or embodied agents but also in the capacity of these systems to answer appropriately to a user’s request and to engage users in a meaningful and adaptable interactive exchange with artificial agents [5, 6]. Chatbot systems might have a diverse knowledge orientation (e.g., generalist or domain-specific) and be designed on the basis of the context of use to have a different range of functionalities and modalities of answering the end-users, for instance, by following algorithms to present their answers or by relying on artificial intelligence and machine learning approaches for retrieving and generating answers to the users

[1]. The diversity of the designed characteristics of the chatbots and their functionalities are determinants of the quality experienced by the end-users. However, as recognized by researchers [9, 14], a common framework to assess the chatbot’s technical performance and user experience is currently missing. As suggested by Jenneboer et al. [22], the system quality perceived by the end-users (i.e., satisfaction) is a key aspect that should be captured to model the overall experience of the users (ISO 9241–210, 2010), and yet how to assess this aspect of the system in a standardized way is unexplored. For this reason, experts are forced to assess the interaction with chatbots using only qualitative and non-comparable observations or by using scales developed for point-and-click interaction that don’t consider aspects such as conversational quality. Acknowledging this issue, Borsci and colleagues [5–7], Borsci, Schmettow, et al., 2022) recently proposed and validated a scale in English, German, Spanish, and Dutch to assess the satisfaction of end-users—i.e., perceived usability (ISO 9241–11, 2018)—after the interaction with a chatbot: the ChatBot Usability Scale (BUS).

The BUS is composed of 11 items (BUS-11) on a 5-point Likert scale from strongly disagree to strongly agree and 5 factors (see Table 1). This inventory is divided into five factors, and it resulted in highly reliable (Cronbach’s $\alpha > 0.9$) and strongly correlated to the UMUX-Lite scale [24]. Currently, this scale is not validated in Italian, and the present study contributes to our understanding both in terms of the adaptation and validation of the Italian version of the BUS-11 based on the original English version (Borsci, Schmettow, et al., 2022) by analyzing reliability, criterion validity, and construct validity on a population of Italian participants. This process could also inform future adaptation of this scale in other languages.

Fig. 1 Conference and peer-reviewed journal articles on Scopus that contains the word chatbots in the title or abstract or keywords from 2012 to 2021

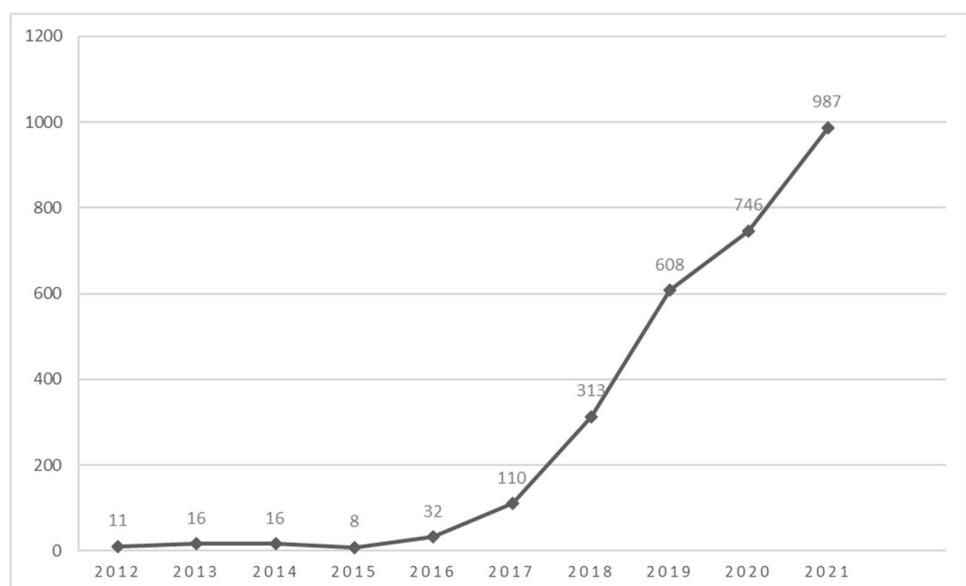


Table 1 Presents the five factors and the eleven items that are composing the Chatbot Usability Scale (BUS-11) as emerged from the original validation [7]

Factors	Item-English version
1-Perceived accessibility to chatbot functions	1. The chatbot function was easily detectable
2-Perceived quality of chatbot functions	2. It was easy to find the chatbot
3-Perceived quality of conversation and information provided	3. Communicating with the chatbot was clear
	4. The chatbot was able to keep track of context
	5. The chatbot's responses were easy to understand
	6. I find that the chatbot understands what I want and helps me achieve my goal
	7. The chatbot gives me the appropriate amount of information
	8. The chatbot only gives me the information I need
	9. I feel like the chatbot's responses were accurate
4-Perceived privacy and security	10. I believe the chatbot informs me of any possible privacy issues
5-Time response	11. My waiting time for a response from the chatbot was short

In line with previous studies on the BUS scale (Borsci, Schmettow, et al., 2022), the Italian version of the scale is expected to have a high level of reliability, e.g., Cronbach's alpha over 0.7, and to be positively correlated with the UMUX-Lite scale [3], i.e., convergent validity. It should also be noted that the BUS questionnaire was previously assessed in a digital form with items presented to participants in a randomized order [5-7], Borsci, Schmettow, et al., 2022). Nevertheless, the scale could also be used as a paper and pencil questionnaire with a predefined order of the items. This aspect was not investigated previously. Therefore, a further goal of this study is to investigate whether the order of the items presented to the participants affects their level of satisfaction with chatbots.

2 Methods

2.1 Study design

The process of translating the scale was performed following the indications from the work of Beaton et al. [2] regarding the cross-cultural adaptation of scales. However, during the translation process and back translation of the BUS-11 in the Italian language, both of the translators independently expressed a similar doubt regarding the first item of the scale demanding a further specification of the question. It was suggested that the first item of the scale—The chatbot function was easily detectable—could be interpreted by Italian end-users as vague or not clear without specifying to which function the item is referring too. In previous studies, item 1 was not highlighted as problematic (Borsci, Schmettow, et al., 2022). However, all the authors agree that to further clarifying this item might potentially help users to better understand the item. Specifically, item 1 is meant to point end-users to

the fact that chatbots, in many cases, are presented as sub-systems within an interface, and that these sub-systems might have their functionalities, such as the possibility to change the appearance of the avatar. Considering the potential issue of comprehensibility for the Italian respondents, it was decided to explore the effect of a rewording of the first item (adaptation of the scale) by presenting Italian participants with two versions of the scale (see Appendix 1: (i) version A: with the first item translated literally and, (ii) ITA BUS version 2, in which item 1 was further clarified; this would read in English as follows:—The chatbot function was easily detectable (e.g., the possibility to modify the settings of the chatbot, make the avatar visible or not.

A cross-sectional between-subjects' design was selected to compare the satisfaction rated by the participants after the interaction with chatbots by using two versions (A and B) of the ITA BUS scale. The results could potentially indicate that one of the two versions (A or B) might show a better fit with the overall construct of the inventory, thus leading to the conclusion that the BUS-11 might benefit from a rewording of item 1 in all the other languages. Participants were randomly assigned to interact and assess a set of four chatbots (out of a list of ten) in one of the two conditions—i.e., the two versions of the scale. Moreover, to test the effect of the order of presentation of the items, participants in each condition (survey version) were further randomized into two groups: one receiving the items of the BUS scale in a randomized order, and one receiving the items in sequential and progressive order.

2.2 Ethics statement

The study was approved by the Ethics committee of the University of Twente (n 211,383), and it was done in collaboration with the University of Modena Reggio Emilia, the University of Pisa, and the University of Perugia.

2.3 Measurements and apparatuses

The study was performed online using a Qualtrics, and each participant performed using their own device (i.e., desktop, tablet, smartphone). The Qualtrics system was used to guide the participants step-by-step and to enable them to receive specific tasks and goals that they were asked to achieve while interacting with the chatbots. Moreover, the online system contains four sections as follows (see Appendix 2 for an overview of the online study). First the participant information and informed consent. Second, the demographics section was composed of questions regarding for instance age, gender, and nationality.

A pool of 10 chatbots was selected for this study (see appendix 3). During the interaction section in Qualtrics, participants (who tested randomly 4 out of the 10 chatbots) received a task specifically associated with the chatbot (see the example in Appendix 2, Sect. 3).

The Italian version of the BUS-11 (See the English version in Table 1, and the two Italian versions (A and B) in Appendix 1) and the Italian Version of the UMUX-Lite composed of 2 items [3] were used for the assessment of satisfaction and criterion validity. The order of presentation of the two scales was randomized.

After the interaction with the four chatbots, participants were asked to comment or report (open question) about issues due to the duration of the study or in understanding the indications and the items.

2.4 Participants

Participants were invited to participate voluntarily in the study by posting announcements on the online systems of different Italian Universities and in different courses from psychology to industrial engineering and computer science. Advertisement of the study was also performed on social media (e.g., Facebook and LinkedIn) to expand the likelihood of gathering insights from participants from different communities and different ages. No specific requirements were needed to participate in the study except for being native to Italian or with a high level of proficiency in the Italian language. A total of 340 participants (218 female, Age mean: 27.31–min 18, max 63) study started and finalized the study. As each participant was asked to interact with and assess four chatbots, a total of 1360 completed observations were collected; nevertheless, the evaluation of one chatbot was excluded as the chatbot was under maintenance and a participant reported that she/he could not access it. Out of the 1359 completed questionnaires, 689 were presented as version A of the scale and 670 as version B.

2.5 Procedure

Participants performed online following the indications on the screen provided by the Qualtrics system. At the beginning of the survey, participants received information about the study and were asked to sign the consent form and answer the questions about the demographics.

After the introduction to the interaction section, participants were asked to interact with the four chatbots. Each chatbot was provided with a task and a link that opened a new tab on the browser, thus enabling participants to perform the task with the application and easily assess their interaction with the BUS-11 and the UMUX-Lite. It was clarified to participants that the objective was to gain enough interaction with the chatbot to assess their satisfaction with the chatbot more than merely fulfilling the task or assessing the possibility to achieve the goal of the task. Participants were instructed to return to the Qualtrics system when they have acquired enough information to assess the quality of the chatbot. At the end of the study, participants were asked to report issues associated with the study regarding, for instance, duration, clarity of the requests, and the items, and participants were thanked for their participation.

2.6 Data analysis

Data analysis was performed in R. Descriptive analysis and linear regression with Scheffé's correction for pair-wise comparison [26] were used to observe differences in users' satisfaction after the interaction with the chatbots. Linear regression model analysis was used to explore the differences in the rating of the respondents in the two conditions (A and B) regarding item 1, Factor 1 (average of Item 1 and Item 2) and the overall score of the BUS-11 scale.

A multi-group confirmatory factorial analysis (MGCFA, [8] approach was applied to explore how the two versions of the scale fit the expected model of factors loading emerged from the validation study of the original version of the scale (Borsci, Schmettow, et al., 2022). The "lavaan" package of R was used for the MGCFA approach. MGCFA was used to validate the ITA BUS and to test if the two versions have differences in terms of factorial structure. MGCFA analysis not only enables the researchers to test the fit of the scale with the expected model but it also allows researchers to test whether an inventory elicits similar response patterns across samples, in our cases the two versions of the scale with different translations of item 1 [8].

Regarding the model fit, the factor loading was considered acceptable when at least 0.6 and optimal at 0.7 and above [16]. Model fit was established by looking at multiple criteria including [11],Hu & Bentler, 1999)the following: 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 standardized root mean square residual (SRMR) looking for a value below 0.08. This confirmatory analysis will inform if the data are fitting the expected model of five factors, as this is a precondition for the MGCFA. The analysis of invariance was performed to observe if the model fit does not change due to the group conditions [12]. A preliminary condition for this test is that the model fit is at least acceptable for the two groups, if this condition is met, the invariance test is performed by adding constraints to the model in an iterative way in four steps, as follows:

- Configural invariance test. This test checks that the factors and pattern of loadings are the same for both groups.
- Metric invariance test. This tests if each item contributes to the latent construct to a similar degree across groups.
- Scalar invariance test. This analysis assesses the mean differences in the latent construct and captures all mean differences in the shared variance of the items.
- Strict invariance test. This is a constrained version of the scalar model where the factor loadings, intercepts, and residual variances are fixed across groups.

A scale is considered invariant if the CFI after each test changes less than 0.01 (ΔCFI , [10]), suggesting that there are no significant effects of the group differences on the factorial structure of the scale.

Comments of the participants at the end of the study regarding issues with the indications and the items were explored to identify if issues emerged, particularly, regarding item 1 of the scale.

Finally, Cronbach Alpha was calculated for the overall scale and per each factor of the BUS-11. The minimum acceptable value for Cronbach’s α was considered to be ≥ 0.70 ($\alpha \geq 0.9 =$ excellent; $0.9 > \alpha \geq 0.8 =$ good; $0.8 > \alpha \geq 0.7 =$ acceptable); below this value, the internal consistency was considered to be low ($0.7 > \alpha \geq 0.6 =$ questionable; $0.6 > \alpha \geq 0.5 =$ poor; $0.5 > \alpha =$ unacceptable) [28]. To establish convergent reliability, a Kendall tau correlation analysis was performed between BUS and UMUX-LITE.

3 Results

3.1 Participants’ satisfaction with the chatbots

Participants were acceptably satisfied by the chatbots (average: 68.9%; SD.1.6%). A linear model analysis suggested that the two versions (A and B) of the scale do not affect the overall scores of the participants regarding their satisfaction with the chatbots. Table 2 summarizes the differences between the average satisfaction per chatbot. An ANOVA

analysis suggests an effect of the type of chatbot on the satisfaction rating of the participants ($F(9, 1349) = 3.656$; $p = 0.0001569$); Scheffe’s test clarified that this difference was significant only for chatbots 10 and 7 ($p = 0.0019$) that differ of 9.9% in terms satisfaction score; indeed, participants rate chatbot 10 as the most satisfactory in the cohort (74%) and chatbot 7 as the less satisfactory one (64.1%).

3.2 Adaptation and validation of the Italian version of the scale

When comparing the response of participants for item 1 and factor 1 in the two conditions, participants who received the new version of the scale (Version B) were slightly, but significantly, more conservative at answering this question compared to participants who answered version A. Specifically, when answering the first item using version B, participants rated lower (-6.3% , CI: -8.55% ; -4.06%) their agreement with item 1 compared to participants who answered version A of the scale ($F(1, 1357) = 30.35$, $p < 001$). A similar effect is also reported on the score of F1 as participants rated lower (-3% ; CI: $-4.9-1.2\%$) compared to the other participants ($F(1, 1357) = 10.85$, $p < 001$). Albeit significant, these effects are quite marginal in terms of scores variation. It should be noted, indeed, that the overall level of satisfaction reported by the two groups answering the BUS-11 was not significantly different, suggesting the two different versions of the scale do not result in different levels of satisfaction.

To model the factorial structure of the Italian BUS-11 and investigate if the different wording of item 1 affected such structure, a MGCFA was performed. The results (see Table 3) suggest that both versions A and B fit the original model with no significant variations in the factorial structure, i.e., the invariance analysis showed no significant changes in the fit ($\Delta CFI < 0.01$) due to the conditions. However, version A appears to have a lower fit (CFI and RSMEA) compared

Table 2 The total number of evaluations collected (observations) per each chatbot, the average score of participants who received different versions of the scale (A and B) and the overall average of the scale

Chatbot	Number of observations	Version A	Version B	Overall average BUS 11
1	142	68.70%	67.10%	67.90%
2	125	70.70%	68.60%	69.65%
3	138	68.00%	72.00%	70.00%
4	135	68.00%	71.70%	69.85%
5	138	67.60%	65.90%	66.75%
6	146	64.90%	73.50%	69.20%
7	138	63.10%	65.30%	64.20%
8	136	66.60%	68.20%	67.40%
9	128	69.40%	69.80%	69.60%
10	133	75.90%	72.40%	74.15%

to version B, suggesting that the adapted version could be considered a better solution in terms of fit quality.

The qualitative analysis of the comments of the participants is quite in line with the results of the MGCFA. Out of 109 participants who decided to leave a comment about the survey, 77% of the comments were positive feedback about instructions, the items on the scales, and the overall experience, and only 23% of the comments (i.e., 25 comments) reported issues experienced by the participants—about five comments were about the length of the study; three about the necessity to provide more instructions to better guide the participants; and 15 comments were about technology-related aspects such as visualization of the chatbots due to device and browser compatibility, or the slow answer of some chatbots compared to others, etc. Finally, only 2 participants (0.05% of the overall sample) commented on the understandability of item 1. Specifically, one comment was about version A, asking for a further explanation about the item, and the other comment was about the same item but in version B, suggesting that not all chatbots present the possibility to change settings.

Considering the results of the MGCFA, the two versions were brought together to model the overall structure of the Italian Version of the BUS-11 as displayed in Fig. 2. The Italian version of the BUS-11 shows a solid factor loading in line with the original validation, despite item 2 (i2, Fig. 2) seems to marginally contribute to factor 1 (fc1, Fig. 2). Moreover, in line with previous validation in English, factors 2 and 3 strongly correlate.

3.3 Convergent validity, reliability, and order of the item presentation

The Italian version of the BUS significantly correlates with the UMUX-Lite ($r_b = 0.67$; $p < 0.001$). Specifically, all the factors of the ITA BUS appear to significantly correlate with the UMUX-Lite overall score with different strengths, as reported in Table 4. Factors 2 and 3 show a good relationship with the UMUX-Lite, while factors 1, 4, and 5 are only marginally correlated with the overall score of the UMUX-Lite.

In line with the expectations, the Cronbach alpha of the overall scale of 0.921 ($F1 \alpha = 0.705$; $F2 = 0.869$, $F3 = 0.903$). Finally, a regression analysis performed on the overall score of satisfaction of the participants who received the scale in a randomized order or a fixed order showed no significant differences in satisfaction scores, suggesting that the scale could be used both presenting the items randomized or in a sequential order without any effect on the assessment of the users.

4 Discussion and conclusion

The Italian version of BUS-11 respects the original factorial structure with high reliability and good convergent validity. The BUS-11 strongly correlates with UMUX-Lite; however, essential aspects such as perceived accessibility (fc1), privacy (fc4), and time to answer (fc5) captured by the BUS-11 are only marginally captured by UMUX-Lite. Moreover, the scale can be used in a digital version with a fully randomized presentation of the items or in a paper–pencil version with items presented in sequential order, without any significant effect on the satisfaction rating. In this sense, the BUS-11 is a compact and reliable scale that, compared to other shorter scales of stratification, could also provide more useful insights into satisfaction with chatbots.

The adaptation of item 1 we proposed seems to change the reaction of participants significantly but slightly without altering the overall scale construct and the overall score of satisfaction, but with a superior fit in terms of constructs. The qualitative analysis we performed on users' reports about issues with the scale seems to suggest the wording of item 1 is a minor problem that affects only a small percentage of the participants, but the proposed adaptation of the item seems to help participants to focus more on identifying the aspects associated with the question at hand resulting in a better fit in terms of construct and it could be interpreted as a positive addition to the scale. Therefore, we can suggest that the adaptation of item 1 could be retained, and future studies should also investigate if this adaptation should be also extended to the English version of the scale.

Table 3 Invariance analysis

	X2 (Df)	RSMEA	SRMR	CFI	Δ CFI
Model					
ITA BUS 11 (all data)	(36) = 269.849	0.069	0.030	0.974	
ITA BUS 11 version A	(36) = 199.813	0.078	0.031	0.968	
ITA BUS 11 version B	(36) = 144.641	0.067	0.032	0.973	
Invariance tests					
Configural invariance	(84) = 383.395	0.072	0.039	0.967	n/a
Metric invariance	(78) = 355.052	0.072	0.036	0.970	0.003 (acceptable)
Scalar invariance	(84) = 383.395	0.072	0.039	0.967	0.003 (acceptable)
Strict invariance	(93) = 417.251	0.072	0.043	0.965	0.002 (acceptable)

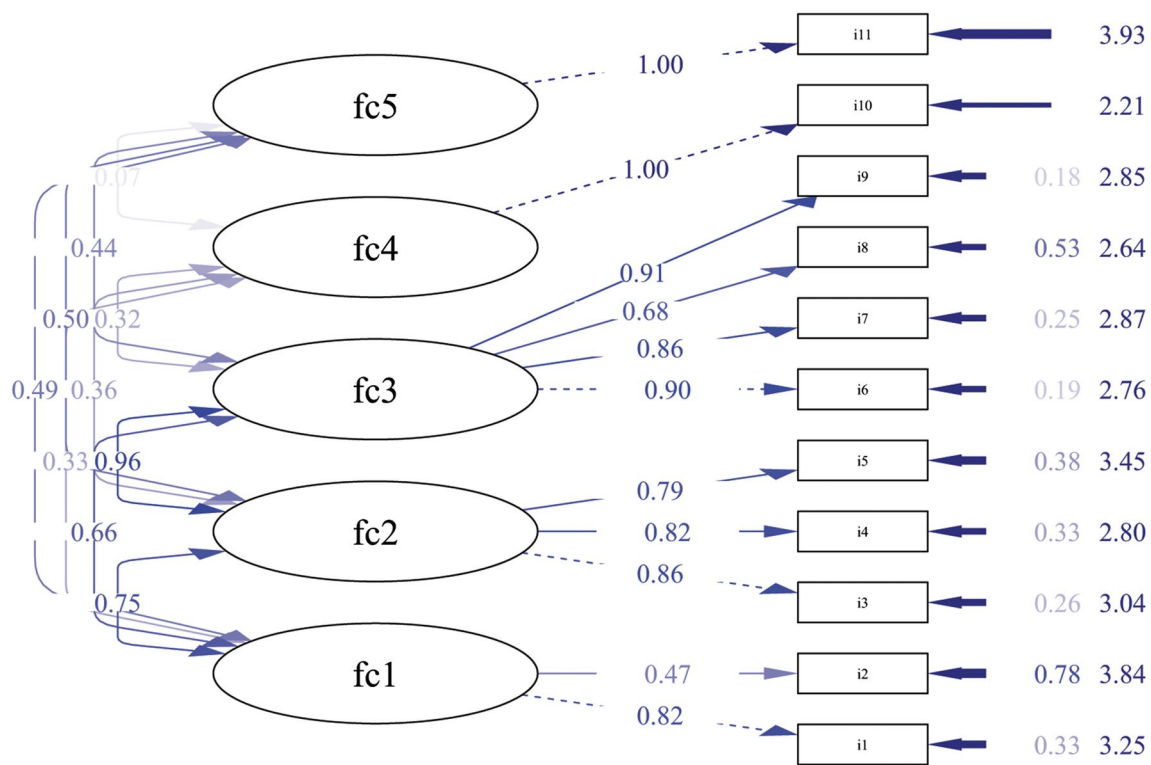


Fig. 2 Graphical presentation of the Chatbot Usability Scale (BUS 11) factorial organization

Table 4 Correlation between ITA BUS 11 and UMUX-Lite

ITA BUS	UMUX Lite overall score
fc1-perceived accessibility to chatbot functions	0.44*
fc2-perceived quality of chatbot functions	0.65*
fc3-perceived quality of conversation and information provided	0.63*
fc4-perceived privacy and security	0.20*
fc5-time response	0.35*
Overall score	0.67*

* $p < 0.001$

The possibility to use a validated scale in Italian will help to shape the future market in the country and to diffuse a culture of benchmarking across companies and countries. This is in line with the idea of Borsci et al. [4] that usability is a construct that needs comparative, standardized, and replicable measures to maximize its effects and enable a positive user experience. The advancement in artificial intelligence research and applications is

pushing chatbots and conversational systems in sensitive and high-risk domains such as education, training, and rehabilitation. This is particularly relevant in Europe, in which evidence of the trustworthiness of products is an essential condition for market access, especially for AI-driven tools [5-7]. In such context, the BUS-11 could be the basis to generate evidence about the perceived quality of AI and algorithmic chatbots, as this scale offers a quick and standardized assessment tool that designers can use to avoid unsatisfactory and potentially problematic solutions to reach the market.

Currently, BUS-11 was mainly tested using commercial and service-oriented chatbots, and future studies should focus on testing the scale with applications for different domains and under different conditions, e.g., using mobile phones or using mainly or only voice exchanges. Moreover, the current version of the scale should be tested in terms of inclusivity, aiming to be accessible and usable by a population of people with different types of disabilities. Future studies are going to explore the possibility to adapt the BUS-11 to enable people with disability to assess their interaction with chatbots.

Appendix 1. The Italian version of the Chatbot Usability Scale (BUS-11): Version A literal translation (ITA BUS A), Version B adapted translation (ITA BUS B)

Factors	Items	
	Italian version (ITA BUS A)	Italian version (ITA BUS B)
1 – Accessibilità percepita alle funzioni del chatbot	Le funzioni del chatbot sono facilmente identificabili	Le funzioni del chatbot sono facilmente identificabili (p. es: come modificare le impostazioni del chatbot, la presenza o meno di un avatar ecc.)
2—Qualità percepita delle funzionalità del chatbot	È stato semplice trovare il chatbot	È stato semplice trovare il chatbot
	La comunicazione con il chatbot è stata chiara	La comunicazione con il chatbot è stata chiara
	Il chatbot è stato capace di tenere conto di quello che ci siamo detti nel contesto della conversazione	Il chatbot è stato capace di tenere conto di quello che ci siamo detti nel contesto della conversazione
3—Qualità percepita delle conversazioni e informazioni fornite	Le risposte del chatbot sono state facili da capire	Le risposte del chatbot sono state facili da capire
	Trovo che il chatbot abbia capito quello che volevo e mi abbia aiutato a raggiungere i miei obiettivi	Trovo che il chatbot abbia capito quello che volevo e mi abbia aiutato a raggiungere i miei obiettivi
	Il chatbot mi ha offerto una quantità adeguata di informazioni	Il chatbot mi ha offerto una quantità adeguata di informazioni
4—Privacy e sicurezza percepite	Il chatbot mi ha dato solamente le informazioni di cui avevo bisogno	Il chatbot mi ha dato solamente le informazioni di cui avevo bisogno
	Penso che le risposte del chatbot siano state accurate	Penso che le risposte del chatbot siano state accurate
	Credo che il chatbot mi abbia informato di tutte le possibili questioni inerenti alla privacy	Credo che il chatbot mi abbia informato di tutte le possibili questioni inerenti alla privacy
5—Tempo di risposta	Il tempo che ho dovuto attendere per le risposte del chatbot è stato breve	Il tempo che ho dovuto attendere per le risposte del chatbot è stato breve

Appendix 2. Overview of the online study

Section 1. Participants' information and informed consent.
Section 2. Demographics.

1. Gender (as assigned at birth)
2. What gender do you identify with? (Please, indicate all relevant options)
3. What is your year of birth
4. What is your nationality?

Section 3. Interaction (example of one chatbot).

In the next section, you will be asked to interact with 4 different chatbots and for each to perform a specific task (for example, find public transport subscription offers) and then answer two questionnaires on the satisfaction of interacting with the chatbot. Specifically: 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 you will have to perform are a way to collect information regarding the functioning of the chatbot 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.iRobot Chatbot

Remember that the purpose is to evaluate the satisfaction of interacting with the chatbot. Regardless of whether or not you complete the task listed below when you think you have acquired enough information to be able to assess the interactive quality of the chatbot you can proceed to fill in the questionnaire.

Task: You are thinking of buying a Roomba vacuum cleaner and you want to know if there are any active offers for one of the models available. Search for this information via the iRobot chatbot.

Go to the website to interact with the iRobot chatbot: <https://www.irobot.it/> < Blank page >

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

Assessment.

Randomized BUS 11 (version A or B)—UMUX-Lite.

[Next Chatbot—Total of 4 chatbots].

Section 4. End of the Survey.

- Open question about issues with the scale: Please report if you have encountered any problems in completing the questionnaire (e.g., duration, clarity of requests and items etc.).
- Final message to thank the participants

Appendix 3. List of chatbots (in alphabetical order)

iRobot—<https://www.irobot.it/>
 PosteItaliane—<https://www.poste.it/chatta-con-noi.html>
 Ryanair—<https://help.ryanair.com/hc/it/articles/360017683957-Italia>
 Solidarietà Veneto—<https://www.solidarietaveneto.it/>
 Tim Italia—<https://www.tim.it/assistenza>
 Tupigi—<https://www.tupigi.it/it/>
 Univerity of Brescia—<https://www.unibs.it/it/didattica/informazioni-utili/contatta-le-segreterie>
 Userbot.ai—<https://userbot.ai/it/>
 Virtual spirit—<https://www.virtualspirits.com/it/>
 Zoom—<https://zoom.us/>

Acknowledgements Dr. Alan Chamberlain's part in this work was supported by the Engineering and Physical Sciences Research Council (grant number EP/V00784X/1) UKRI Trustworthy Autonomous Systems Hub and (grant number EP/S035362/1) PETRAS 2.

Declarations

Competing interests The authors declare no competing interests.

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

- Adamopoulou E, Moussiades L (2020) Chatbots: history, technology, and applications. *Machine Learning with Appl* 2:100006. <https://doi.org/10.1016/j.mlwa.2020.100006>
- Beaton DE, Bombardier C, Guillemin F, Ferraz MB (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25(24):3186–3191. <https://doi.org/10.1097/00007632-200012150-00014>
- 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-Computer Interact* 31(8):484–495. <https://doi.org/10.1080/10447318.2015.1064648>
- 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>
- Borsci S, Lehtola VV, Nex F, Yang MY, Augustijn E-W, Bagheriye L, Brune C, Kounadi O, Li J, Moreira J, Van Der Nagel J, Veldkamp B, Le DV, Wang M, Wijnhoven F, Wolterink JM, Zurita-Milla R (2022) Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle. *AI & Soc*. <https://doi.org/10.1007/s00146-021-01383-x>
- Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, Chamberlain A (2022) The Chatbot Usability Scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. *Pers Ubiquit Comput* 26(1):95–119. <https://doi.org/10.1007/s00779-021-01582-9>
- Borsci S, Schmettow M, Malizia A et al (2023) A confirmatory factorial analysis of the chatbot usability scale: a multilanguage validation. *Pers Ubiquit Comput* 27:317–330. <https://doi.org/10.1007/s00779-022-01690-0>
- Brown GTL, Harris LR, O'Quin C, Lane KE (2017) Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education* 40(1):66–90. <https://doi.org/10.1080/1743727X.2015.1070823>
- Caldarini G, Jaf S, McGarry K (2022) A literature survey of recent advances in chatbots. *Information* 13(1):41. <https://doi.org/10.3390/info13010041>
- Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct Equ Model* 9(2):233–255. https://doi.org/10.1207/s15328007sem0902_5
- Cole DA (1987) Utility of confirmatory factor analysis in test validation research. *J Consult Clin Psychol* 55(4):584
- D'Urso ED, De Roover K, Vermut JK, Tijnstra J (2021) Scale length does matter: recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. *Behav Res Methods*. <https://doi.org/10.3758/s13428-021-01690-7>
- Saracco R (2022) Perspectives on AI adoption in Italy, the role of the Italian AI Strategy. *Discov Artif Intell* 2:9. <https://doi.org/10.1007/s44163-022-00025-5>
- Federici S, de Filippis ML, Mele ML, 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. *Disabil Rehabil: Assist Technol* 15(7):832–837. <https://doi.org/10.1080/17483107.2020.1775313>
- Gupta S, Chen Y (2022) Supporting inclusive learning using chatbots? A chatbot-led interview study. *J Inf Syst Educ* 33(1):98–108
- Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2014) *Pearson new international edition. Multivariate data analysis, 7th Edn.* Pearson Education Limited Harlow, Essex
- He L, Basar E, Wiers RW, Antheunis ML, Kraemer E (2022) Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health* 22(1):726. <https://doi.org/10.1186/s12889-022-13115-x>
- Hollebeek LD, Glynn MS, Brodie RJ (2014) Consumer brand engagement in social media: conceptualization, scale development and validation. *J Interact Mark* 28(2):149–165. <https://doi.org/10.1016/j.intmar.2013.12.002>
- Hu L, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Modeling* 6(1):1–55
- ISO (2010) *ISO 9241–210:2010 Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems.* Brussels, BE: CEN
- ISO (2018) *ISO 9241–11 Ergonomic requirements for office work with visual display terminals—Part 11: Guidance on usability.* Brussels, BE: CEN
- Jenneboer L, Herrando C, Constantinides E (2022) The impact of chatbots on customer loyalty: A systematic literature review. *J*

- Theor Appl Electron Commer Res 17(1):212–229. <https://doi.org/10.3390/jtaer17010011>
23. Jiang H, Cheng Y, Yang J, Gao S (2022) AI-powered chatbot communication with customers: dialogic interactions, satisfaction, engagement, and customer behavior. *Computers Human Behav* 134:107329. <https://doi.org/10.1016/j.chb.2022.107329>
 24. Lewis JR, Utesch BS, Maher DE (2013) UMUX-LITE: when there's no time for the SUS proceedings of the SIGCHI conference on human factors in computing systems, Paris. France DOI 10(1145/2470654):2481287
 25. ReportLinker (2021) North America and Europe chatbot market-growth, trends, COVID-19 impact, and forecasts (2021–2026). Retrieved July 2 from: <https://www.globenewswire.com/news-release/2021/10/27/2321521/0/en/North-America-and-Europe-Chatbot-Market-Growth-Trends-COVID-19-Impact-and-Forecasts-2021-2026.html>
 26. Scheffé H (1953) A method for judging all contrasts in the analysis of variance. *Biometrika* 40(1–2):87–110
 27. Spagnolo E (2022) The artificial intelligence market is growing in Italy. Retrieved July 2 from: <https://en.cryptonist.ch/2022/02/07/the-artificial-intelligence-market-is-growing-in-italy/>
 28. Tavakol M, Dennick R (2011) Making sense of Cronbach's alpha. *Int J Med Educ* 2:53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
 29. Verhoef PC, Neslin SA, Vroomen B (2007) Multichannel customer management: understanding the research-shopper phenomenon. *Int J Res Mark* 24(2):129–148. <https://doi.org/10.1016/j.ijresmar.2006.11.002>

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