Unleashing the potential of Turkish chatbots
a study on the validity and reliability of the bot usability scale

Aktaş, Neşe Baz; Şişman, Burak; Borsci, Simone

DOI
10.1007/s10209-025-01211-9

Publication date
2025

Document Version
Final published version

Published in
Universal Access in the Information Society

Citation (APA)
Aktaş, N. B., Şişman, B., & Borsci, S. (2025). Unleashing the potential of Turkish chatbots: a study on the validity and reliability of the bot usability scale. *Universal Access in the Information Society*, *24*(3), 2467-2476. https://doi.org/10.1007/s10209-025-01211-9

Universal Access in the Information Society
https://doi.org/10.1007/s10209-025-01211-9

LONG PAPER

Unleashing the potential of Turkish chatbots: a study on the validity and reliability of the bot usability scale

Neşe Baz Aktaş[1,4] · Burak Şişman[2] · Simone Borsci[3,5]

Accepted: 26 February 2025

Abstract
The objective of this study is to adapt and evaluate the Turkish version of the Chatbot Usability Scale (BUS-11) through a confirmatory factor analysis method. The BUS-11 scale has been established in various languages except for Turkish; thus, its validation and dissemination could serve as a means to improve chatbot interaction satisfaction among the Turkish-speaking population and hence foster growth in Turkey's conversational agent market. To achieve this aim, seven customer-oriented chatbots were rated on pre-designed tasks by participants. Data collection involved using the Turkish-adapted BUS11 (TBUS-11) to assess individuals' experiences after interacting with Turkish-speaking chatbots, along with the Turkish version of the UMUX-LITE scale. Results show that TBUS-11 has been demonstrated to be highly reliable with a strong convergent validity with the UMUX-LITE already validated in Turkish. Moreover, the analysis demonstrated that the dataset supported the five-factor structure of the original version of the scale, thus confirming the psychometric properties of the TBUS. The study successfully adapted the BUS-11 into Turkish, providing a reliable and valid tool for assessing chatbot usability in the Turkish-speaking market. This can potentially enhance user satisfaction and promote the growth of conversational agents in Turkiye.

Keywords Chatbot · Scale adaptation · Interaction satisfaction · Usability · User experience · Human-computer interaction

1 Introduction

Chatbot usage has gained widespread popularity in recent years, particularly for business and customer service purposes [12, 34]. In many fields, including healthcare, finance, and e-commerce, chatbots have become a favored tool for interacting with consumers. Chatbots employ natural language for engaging in effective interactions with users and providing responses to their inquiries. Currently, chatbots are utilized in various domains to offer round-the-clock services. Chatbots possess the ability to address a minimum of 80% of typical customer queries and are accessible round-the-clock, indicating their versatility beyond just being limited to handling inquiries [1]. Primarily employed for sales-related activities, followed by support and marketing functions, chatbots have demonstrated a notable impact on increasing sales figures with an average improvement of 67% [27]. Additionally, one-fourth of all transactions involve interactions mediated through these automated conversational agents [27]. However, in order to fully leverage the potential of chatbots, it is crucial that their usability be

✉ Neşe Baz Aktaş
nese.aktas1@ogr.sakarya.edu.tr

1 Management Information Systems, Sakarya University, Esentepe Campus, Business Faculty, Sakarya 54050, Turkey

2 TU Delft's Faculty of Mechanical, Maritime and Materials Engineering, 34.H-4-340, Delft, Netherlands

3 Department of Learning, Data analysis, and Technology, Cognition, Data and Education (CODE) Group, Faculty of Behavioural Management and Social sciences, University of Twente, P.O. Box 217, Enschede 7500 AE, Netherlands

4 Department of Industrial Engineering and Business Information Systems, University of Twente, Enschede, The Netherlands

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

Published online: 08 March 2025

evaluated systematically. In other words, the effectiveness and usability of these chatbots remain a subject of inquiry.

On the other hand, the evaluation of user satisfaction with chatbots has become an important area of research to understand how effectively such agents might meet the needs of users enabling a positive experience [18]. Evaluation scales are available and widely used to assess the interaction quality perceived by the users of a digital system, for instance, reliable and concise scales such as the *System Usability Scale* (SUS) [8], the *Usability Metric for User Experience* (UMUX) [11], and UMUX-LITE [21]. The SUS is a "*quick-and-dirty*" questionnaire that consists of ten items using a five-point scale, which has been shown to have excellent psychometric properties. In contrast, UMUX employs fewer items in line with the International Organization for Standardization (ISO) definition of usability and utilizes a seven-point scale ranging from strongly disagree to strongly agree. UMUX Lite is a two-item instrument that has acceptable reliability, validity, and psychometric properties, making it a valuable tool for preliminary and rapid testing of user reactions to a prototype [4]. Nevertheless, while these tools have been widely used and validated for assessing user satisfaction with web interfaces when it comes to chatbots these scales may not fully capture the conversational aspects of user interaction with chatbots since they were not designed for this purpose (Borsci, Malizia, et al., [5]. In a similar fashion, as suggested by Lewis [19, 20], scales to assess the experience with vocally controlled interfaces were developed in the past, for instance, the Mean Opinion Scale [29], the Subjective Assessment of Speech System Interfaces is composed by [15] and the Speech User Interface Service Quality developed by [26]. Nevertheless, voice-controlled interfaces are not comparable to chatbot systems.

The development of a novel measurement scale, known as the Chatbot Usability Scale (BUS-11), is one example of recent efforts to create an evaluation tool that can be used specifically for gauging users' satisfaction following interactions with chatbots. This newly devised scale has been validated and established as fit-for-purpose according to research conducted by Borsci et al. [5]. The BUS-11 scale, composed of 11 items, is structured into five distinct factors: (1) Perceived Accessibility to Chatbot Functions, (2) Perceived Quality of Chatbot Functions, (3) Perceived Quality of Conversation and Information Provided, (4) Perceived Privacy and Security, and (5) Time Response. These factors collectively assess aspects ranging from ease of access and functionality to privacy, security, and responsiveness of chatbots. The scale has been extensively validated and is currently available and validated in Dutch, Italian, German, and Spanish. The researchers utilized a comprehensive approach to develop the construct of such a scale [5]. Additionally, the researchers utilized the UMUX-LITE scale [21] to establish convergent validity as a dependent variable. The outcomes of previous studies suggested a positive correlation between BUS-11 and UMUX-LITE [5]. BUS-11 can aid practitioners in comparing and benchmarking their chatbots during product evaluation, allowing them to enhance their performance. The researchers also suggested that the construct and the factors behind the BUS scale can be also utilized as an efficient checklist for chatbot designers during the development process (Borsci, Malizia, et al., [5].

Turkey has become a desirable e-commerce market to invest in by Turkish companies and foreign partnerships due to its population of over 80 million and the accelerating e-commerce volume [2]. Along with the e-commerce market, Turkey is also a potential growth market for chatbots, with many businesses incorporating chatbots into their operations. CBOT, the leading chatbot developer in the country, claims that 90% of customer inquiries can be handled by customized chatbots with appropriate task customization [17]. However, it is essential to ensure the evaluation metrics used for chatbots are not only specific but also valid and reliable. It is currently not possible to find a validated chatbot usability scale in the Turkish language. Therefore, further research in this area is highly necessary to ensure the efficacy and comprehensiveness of any evaluation tool used for Turkish chatbots. The validation in Turkey of BUS-11 to facilitate the assessment and the comparability of the quality of interaction with chatbots might help practitioners in Turkey to benchmark their chatbots within the country and at international level, this might potentially unify and ameliorate the user experience offered by these systems and thereby improve the performance of chatbots. Therefore, this study aims to adapt the BUS-11 scale for the Turkish language to assess its validity and reliability in evaluating chatbots in Turkey.

2 Method

This study was conducted with the approval of the Istanbul Bilgi University Ethics Committee to ensure adherence to ethical research standards and protect the rights and well-being of the study participants.

2.1 Participants

The study comprised a sample of 199 participants (49% female), with a total of 557 evaluations provided for the chatbots. The study design entails participants engaging with and evaluating a selection of up to four chatbots out of a pool of seven using the Qualtrics Survey platform (see Appendix A for chatbots). The evaluation distribution was as follows: 100 participants assessed four chatbots, 55

participants assessed one chatbot, 30 participants assessed two chatbots, and 14 participants assessed three chatbots. Demographically, the majority (%75.38) of the group belonged to the age bracket between 18 and 24 years old, while there were around 39 individuals (%19.60) aged from 25 to 34 years old and 10 individuals (%5.02) of 35 to 44 years of age. The nationalities represented in this study included 13 participants hailing from Serbia, Romania, Afghanistan, and Azerbaijan. To ensure the results were accurate and reliable, a strict process was followed wherein only the valid responses from participants remained for further analysis.

2.2 Instruments and study procedure

To adapt the scale to Turkish, approval was sought from the authors who created the BUS-11 scale and content validity was considered by consulting with two field experts and one assessment and evaluation expert. The adaptation study aimed at minimizing cultural differences. Language validity of the scale was ensured through a back-translation design aided by two proficient translators who are fluent in both cultures. One of the translators translated from English, the original language of the scale, into Turkish, and the other translated the first translator's translation into English without seeing the original text. Subject matter specialists were consulted in ensuring linguistic equivalence between versions to ensure robust results that minimize sources error when adapting assessment instruments across languages. Additionally, three experts, including the developer of BUS-11, agreed on the quality of the back translation.

The translated scale form was applied to a group of 27 participants (15 female, 12 male) in order to determine the items that were not understood and finalized as a result of the feedback received from the participants. All participants were undergraduate students in Turkish universities. In the pilot study, participants were assigned a task to convert their airline tickets into open ones using an airline chatbot available at the website of an airways company. Following this interaction, they rated their level of satisfaction with the chatbot by answering questions. Regarding understandability, participants were asked to assess the wording of the items on a scale of 1 (not understandable) to 5 (very understandable). The validity of these responses was measured through Cronbach's Alpha which yielded an acceptable score of 0.76. Descriptive statistics were given in Table 1 for the pilot study. Overall, the mean score of each item was high, suggesting that participants had the impression to understand clearly the items. However, item 1 had the lowest mean score and highest standard deviation in terms of understandability. To reduce the risk of potential miscomprehension due to the wording of the translated items, a Turkish language teacher was consulted, who suggested a revision regarding items 1 and 8, so that TBUS-11 was finalized (see Appendix B).

Table 1 Descriptive statistics of the TBUS-11pilot study

TBUS Items	Usability of chatbot per item		Understandability assessment	
	M	SD	M	SD
Item1	3.48	1.28	4.07	1.20
Item2	4.04	1.09	4.78	0.57
Item3	4.19	0.73	4.74	0.59
Item4	4.26	0.85	4.59	0.84
Item5	4.26	0.81	4.67	0.67
Item6	3.78	1.21	4.70	0.60
Item7	4.04	0.98	4.59	0.79
Item8	3.70	1.23	4.52	0.84
Item9	4.22	0.84	4.67	0.67
Item10	4.00	0.87	4.41	0.88
Item11	4.63	0.92	4.78	0.57

In the current study, websites with chatbots in various fields (banking, e-commerce, etc.) that speak Turkish and can be accessed directly (no membership required, etc.) were listed by the researchers. All chatbots were analyzed to avoid selecting chatbots developed by the same developers. In other words, chatbots with the same infrastructure, even if adapted for different sectors, were not selected to ensure diversity. Seven chatbots were identified. Scenarios and tasks were developed for each chatbot to enable participants to use the chatbot for a purpose and elicit particular information (see Appendix A). A two-part survey was designed. In the first part, demographic information was asked to participants. In the second part, participants were asked to choose one of the seven chatbots. They were given a scenario and task developed according to the chatbot they chose and asked to experience the chatbot with this scenario and task. Then, as a form of control, participants were asked to confirm whether they had completed the task and, even if they did not achieve the goal, they were asked whether they had an experience with the chatbot with which they could express their satisfaction. If participants answered positively, they were allowed to answer the items on the TBUS-11 scale and the Turkish version of the UMUX-LITE scale [3, 21]. In addition, as a control question, the participants were asked to say "goodbye" to the chatbot at the end of each task, and then the participants were asked about the chatbot's response to "goodbye". Thus, in terms of validity, it was aimed to determine whether the participants used the chatbot or not. Participants are allowed to choose up to four chatbots in a session.

Pilot study task You purchased a discounted economy-class flight ticket from Pegasus Airlines from Istanbul to Amsterdam. However, due to incomplete visa procedures, you wish to have the ticket converted to an open one. Try using the

chatbot on the website of an airways company to find out when is the latest date you can use your open ticket and how much deduction there will be when you convert it to an open ticket.

2.3 Data analysis

Descriptive and regression analyses were used to assess users' satisfaction after the interaction with all the chatbots. The control of participant interaction was tested by a regression analysis to check the level of satisfaction (TBUS-11) declared by participants who answered correctly or incorrectly the control questions. If participants decided to skip the question, we counted the answer as incorrect. Moreover, a regression was used to test if participants who did not manage to finish the task resulted in less satisfied compared to the overall population.

The data was analyzed using R. After checking for the normality of the distribution by a Shapiro-Wilk test, a confirmatory factor analysis (CFA) was performed using the structural equation modeling function of the 'Lavaan' package. Moreover, an appropriate estimator for non-normally distributed ordinal data was selected between a diagonally weighted least squares (DWLS) or an unweighted least squares (ULS) estimator as the literature suggested that these two are more reliable estimators than classic Maximum Likelihood Robust (MLR) approach for our type of data [22, 25].

Regarding the model fit the factor loading was considered acceptable at a score of at least 0.6 and optimal at 0.7 and above [14]. Model fit was established by looking at multiple criteria including [9, 16]: the ratio between Chi-Square and 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) aimed for values less than 0.07; the Standardized Root Mean Square Residual (SRMR) aimed for a value below 0.08. The R package 'ggplot' was used to visualize the overall model. Cronbach's Alpha was calculated for the overall scale and per each factor of the BUS 11. The minimum acceptable value for Cronbach's α is 0.70, with values below this considered to indicate low internal consistency [33].

To establish convergent validity a Kendall tau correlation analysis was performed between BUS and UMUX-LITE. Finally, as suggested by previous validation studies in other languages [5], we performed analysis of variance and t-test to explore the potential effect of age and sex declared by the participants on satisfaction with chatbots.

Table 2 Descriptive statistics for each chatbot

Chatbot	Number of observations	TBUS		UMUX-LITE	
		M	SD	M	SD
1	70	80.2%	13.9%	80.1%	20.5%
2	86	79.3%	14.4%	80.1%	18.1%
3	45	77.3%	16.6%	77.5%	21.8%
4	72	70.4%	17.9%	69.8%	18.8%
5	95	74.5%	14.4%	76.3%	16%
6	128	78.2%	16.3%	79.3%	17.9%
7	61	77.8%	14.8%	81.3%	17.1%

All scores are presented as percentages, calculated by averaging Likert-scale responses (1 to 5) and converting these averages to percentages to illustrate the proportion of maximum satisfaction

3 Findings

3.1 Control of participants' interaction

Out of 557 observations, in 75 cases the users decided to not answer the control questions, in 242 cases participants answered correctly, and in 240 cases they answered incorrectly. We decided to aggregate missed and incorrect answers. Thus, only 43.4% of the participants demonstrated a certain level of attention to the conversational exchange with chatbots until the end of the conversation. Despite that, the regression analysis suggests no significant differences between people who answered correctly or not correctly to the control questions in terms of satisfaction. This might indicate that even if participants did not remember how the chatbots presented their greetings, the users were able to reliably assess their experience. Therefore, we decided to keep all the observations. However, when performing the CFA we decided to also control the fit with the expected model separately for people who answered correctly or incorrectly to the check questions.

For 55 participants it was observed that they failed to achieve the goal but felt that they could assess the quality of the chatbots anyway. In these cases, the aggregate average satisfaction of participants was 68.6% (SD: 15.5%) compared to the average of the remaining 499 which was 77.8% (SD: 15.4%). Despite the 9.2% difference, a regression analysis indicated no significant difference between the satisfaction of people who achieved or failed the task with the chatbot i.e., $F(1,555) = 1.95$, $p = .163$.

3.2 Satisfaction of participants per each chatbot

Descriptive statistics of satisfaction of participants per chatbot are given in Table 2. The participants' satisfaction with the chatbot interactions, as evaluated using the TBUS and UMUX-LITE, was generally high. According to the findings in Table 2, Chatbot 1 received the highest level of satisfaction (TBUS = 80.2%, UMUX-LITE = 80.14%), whereas Chatbot 4 had a lower satisfaction rating overall (TBUS = 70.4%, UMUX-LITE = 69.86%). The regression analysis clarified that there is a significant effect on

the satisfaction rated by participants due to the chatbots ($F(6,550)=3.5$; $p=.002$). More specifically, compared to Chatbot 1, both Chatbot 5 ($p=.04$) and Chatbot 4 ($p<.001$) were ranked significantly less satisfactory by users in their interactions with them.

3.3 Confirmatory factor analysis of the Turkish adapted version of BUS

The expected five factors of BUS-11 identified in previous validation studies (Borsci, Prati, et al., [6], Borsci, Schmettow, et al., [7] were confirmed in the present experiment (see Table 3, All datasets) in particular the dataset resulted very solid even when participants incorrectly answered the check questions (Table 3, incorrect dataset). The best fit was of course obtained by considering the 242 observations of the participants who correctly answered the control questions (Table 3, correct dataset).

Despite the fit of the correct dataset being certainly superior, it is reassuring that also in the case of participants who gave minimal attention to the task (incorrect dataset), the scale remains solid and the fit with the data is excellent. Figure 1 represents all datasets showing the relationship between items and factors. In line with previous studies a correlation of 0.940 between Factor 2 (Perceived quality of chatbot functions) and Factor 3 (Perceived quality of conversation and information provided).

Table 3 Fit measures of the data from the confirmatory factor analysis*

Dataset	ChiSquare ratio	CFI	RMSEA	SRMR
All datasets	2.14	0.996	0.045	0.040
Incorrect dataset	1.38	0.997	0.035	0.049
Correct dataset	1.07	0.999	0.017	0.039

*Three datasets were tested: The first one is all data, the second one is only data of participants who incorrectly answered the control questions and the third one is the data of only those participants who answered correctly to the control questions

3.4 Reliability, convergent validity of the adapted scale, and effects of individual characteristics

The scale seems to be very reliable with a Cronbach's Alpha of 0.915. Moreover, in line with validation studies in other languages, the TBUS is strongly correlated to UMUX-LITE ($Rt=0.69$, $p<.001$, see Fig. 2).

Users' characteristics do not seem to affect the level of satisfaction measured by TBUS. Specifically, despite a trend (see Fig. 3) the age of users does not seem to affect their rating of satisfaction in our population as suggested by a Kruskal-Wallis's test, i.e., $H(2, n=55)=4.86$, $p=.088$. Concurrently, a t-test analysis indicated that the sex declared by

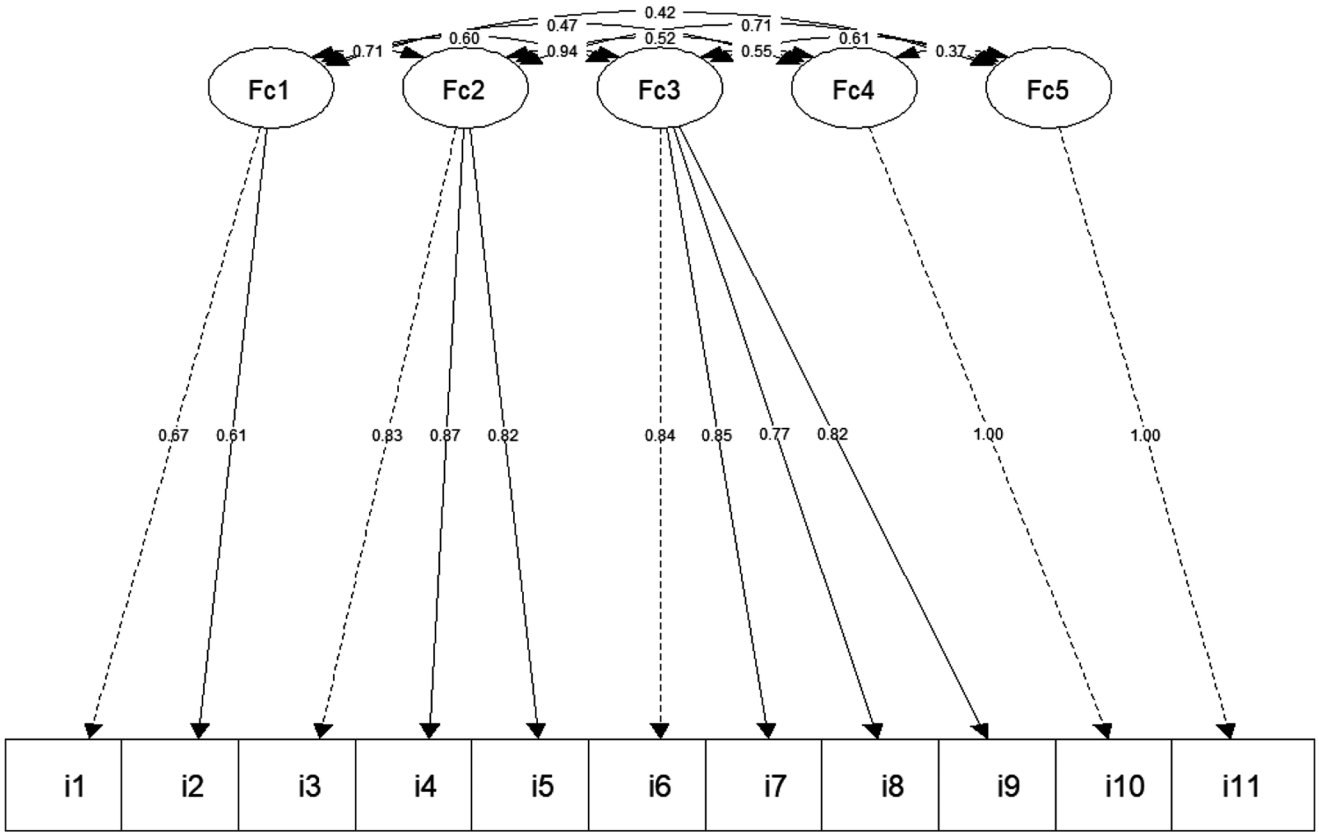

Fig. 1 Graphical presentation of the relationships between factors and items of TBUS-11

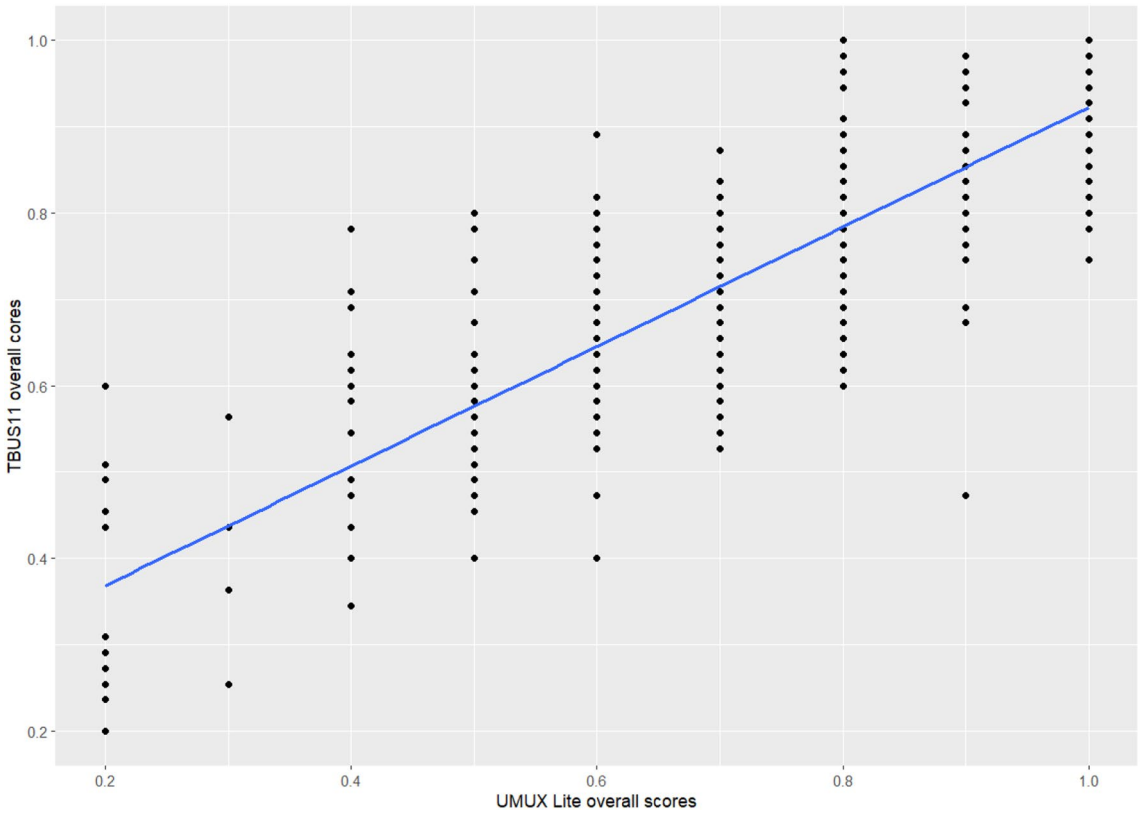

Fig. 2 Correlation between TBUS-11 and UMUX-LITE scale

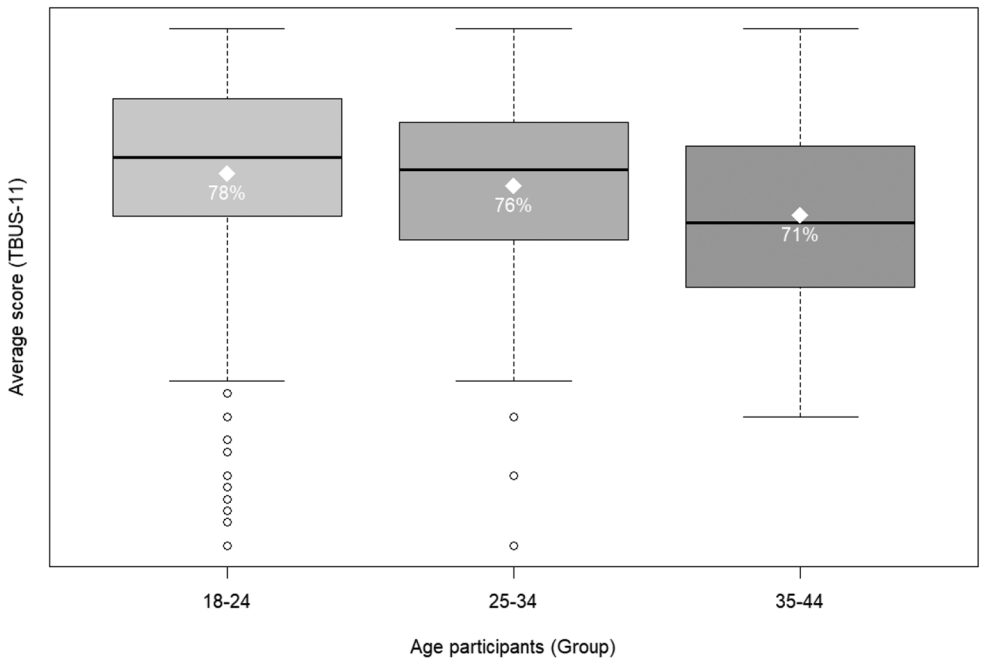

Fig. 3 Relationship between the age groups of participants and TBUS-11 score

respondents does not affect their satisfaction level with the chatbots ($t(558)=2.5$, $p=.016$).

4 Discussion

Based on the TBUS scores, participants perceived the chatbots tested to be satisfactory with an average score ranging from 70% to a maximum of 80.2%. The TBUS-11 scale

aligns with our expected results and well-coincides with the five original BUS-11 structures. Consequently, it provides Turkish users a reliable and standardized way of assessing their satisfaction level when interacting with chatbots. The TBUS scale aims to assess the different aspects that contribute to the conversational system's quality, including Accessibility (F1, average loading of items: 0.64), Perceived quality of chatbot functions (F2, average loading of items: 0.84), Perceived quality of conversation and information providing (F3, average loading of items: 0.82), Perceived privacy and security (F4), and Time Response (F5). F4 and F5 were composed of single items as determined in the original study. In line with findings from previous validation studies (Borsci, Prati, et al., [6], Borsci, Schmettow, et al., [5], there exists a robust correlation between Factor 2 and 3 on the TBUS scale. This indicates that the participants had difficulty distinguishing between factors related to functionality quality (items 3–5) and those associated with conversation and information quality (items 6–9).

The analysis suggested that the Turkish version of the scale is reliable even when participants misbehave or miss to achieve the expected goal of the task for any reason. Specifically, the TBUS analysis showed that these participants reported a less satisfactory experience on average than those who paid attention and successfully completed the tasks. Moreover, the dataset supports the expected five-factor model structure, demonstrating the robustness of the scale. This finding is significant as it suggests that insights derived from TBUS-11 can still be considered reliable even when users do not achieve the intended outcomes for various reasons.

5 Conclusion

This study successfully proposed and validated the Turkish adaptation of the BUS-11 scale, offering a reliable tool to assess user satisfaction with chatbots in tasks related to information retrieval. Our research exemplifies the increasing interest in identifying effective methods to evaluate the quality of human-chatbot interactions [24, 28]. By contributing a standardized assessment tool, this work supports further research and development in the field of conversational agents, particularly within the Turkish-speaking context.

Our goal was to contribute to the domain of conversational agents in Turkey and to facilitate the implementation of high-quality Turkish chatbots by providing a relatively easy-to-use benchmarking tool. The TBUS-11 scale is a valid and reliable tool for measuring user satisfaction with chatbots. The scale is effective in capturing the essential aspects of satisfaction with chatbots and it is perfectly in line with the other versions of the BUS scale validated in other languages. The possibility to access a multilanguage scale such as the TBUS could potentially allow operators in academia and industry to assess their products against national and international systems. This may also in the long run feed cross-cultural comparisons of usage and satisfaction with conversational systems.

Future studies should explore additional factors that influence user satisfaction with chatbots, particularly for users who aim to use the chatbot but are unable to retrieve information or accomplish their intended tasks. The present study also did not investigate other potential aspects affect people's experience with chatbots such as design patterns, trustworthiness, etc [10, 30, 31]. Although our current study did not detect a significant impact of age on satisfaction, it is important to recognize that the participant pool consisted primarily of cooperative undergraduate students who voluntarily participated. Previous research identified an effect of age on participants' answers to the BUS scale, therefore future studies should compensate for that and also test satisfaction with a cohort more representative of people from different ages. In fact, technology usage can vary across different age groups in Turkey, and literature suggests that elderly people in Turkey may have different needs and expectations when using technological artifacts [13]. It is recommended to involve participants from different age groups and backgrounds in order to obtain a more representative sample of the population. Moreover, given the fast-paced changes in technology and user expectations, continued research must be conducted to update the scale as needed to ensure its relevance.

Future studies should also consider the role of chatbot gender presentation, including the use of avatars, in shaping user satisfaction and usability perceptions. Scholars have depicted that the gender of a chatbot can influence how users interact with and evaluate technology [23]: Seo [32],. Exploring the impact of different gender presentations and avatars on user experience could provide valuable insights into user preferences and biases.

Appendix

Table A Tasks and list of chatbots

Chatbot	Tasks
1. Fenerbahçe	"You want to surprise your friend who is a Fenerbahçe fan. You heard that Fenerbahçe offers stadium and museum tours. You can either buy a museum tour or get a gift from the Fenerium store. Use the chatbot at https://www.fenerbahce.org/ to learn about the stadium and museum tour options for the surprise you can make for your friend. Also, find out the location of a Fenerbahçe store you can visit."

Table A Tasks and list of chatbots

Chatbot	Tasks
2. Akbank	"You want to open a bank account for your new job. You cannot leave the office during working hours, but you can visit a bank branch during your lunch break. Use the chatbot at https://www.akbank.com to find the nearest branch that operates during lunch hours and obtain its address."
3. Axa Sigorta	"You will be going to Germany for a one-week vacation and want to get travel insurance from AXA Insurance. Use the chatbot at https://www.axasigorta.com.tr to learn about the details of their international travel insurance. Additionally, find out about the healthcare institutions that AXA Insurance has agreements with."
4. Bağcılar Manicipulity	"You have decided to get married and need to complete the marriage procedures at Bağcılar Municipality. However, you don't know what you need to do. Visit http://www.bagcilar.bel.tr/ and use the chatbot to learn about the required documents, marriage fees, and wedding hall options for your marriage. Also, find out the daily garbage collection times for a friend living on Hürriyet Street in Kirazlı neighborhood."
5. Flormar	"You want to buy a gift for a friend. You learned that your friend needs bronzing powder, eye concealer, and different colored nail polishes. Use the chatbot at https://www.flormar.com.tr/ to find out the prices of these product options and the shipping cost for your friend."
6. Pegasus	"You purchased a plane ticket from Pegasus Airlines for your international trip. Before preparing your suitcase, use the chatbot at https://www.flypgs.com/ to learn the cabin baggage size allowed for your flight."
7. Yapikredi	"You live in Turkey and will be traveling to the Netherlands next week. During your trip, you want to be able to use your Yapıkredi bank card at payment points and ATMs abroad. Use the chatbot at https://www.yapikredi.com.tr to learn about the procedures related to using your card internationally."

Table B Turkish version of BUS-11 and UMUX-LITE

Factors	Items number	Items in English	Items in Turkish
1. Perceived accessibility to chatbot functions	1	The chatbot function was easily detectable, e.g., the possibility to modify the settings of the chatbot, make the avatar visible or not, etc.	Sohbet robotu kolayca fark edilebilirdi. (örneğin; sohbet robotunun ayarlarını değiştirme olanağı, avatarı görünür ya da görünmez yapma tercihleri)
	2	It was easy to find the chatbot.	Sohbet robotunu bulmak kolaydı.
2. Perceived quality of chatbot functions	3	Communicating with the chatbot was clear.	Sohbet robotu ile iletişim anlaşılırdı.
	4	The chatbot was able to keep track of context.	Sohbet robotu konuyu takip edebildi.
	5	The chatbot's responses were easy to understand.	Sohbet robotunun yanıtlarını anlamak kolaydı.
3. Perceived quality of conversation and information providing	6	I find that the chatbot understands what I want and helps me achieve my goal.	Sohbet robotu ne istediğimi anlayabilir ve amacıma ulaşmamda bana yardımcı olur.
	7	The chatbot gives me the appropriate amount of information.	Sohbet robotu bana yeterli miktarda bilgi verir.
	8	The chatbot only gives me the information I need.	Sohbet robotu bana yalnızca ihtiyacım olan bilgileri verir.
	9	I feel like the chatbot's responses were accurate.	Sohbet robotunun yanıtlarının doğru olduğuna inanıyorum.
4. Perceived privacy and security	10	I believe the chatbot informs me of any possible privacy issues.	Sohbet robotunun gizliliği ihlal eden (örneğin; Kişisel Verileri Koruma Kanunu (KVKK) kapsamına giren) herhangi bir durumda beni bilgilendireceğine inanıyorum.
5. Time Response	11	My waiting time for a response from the chatbot was short.	Sohbet robotundan yanıt bekleme sürem kısaydı.

UMUX-LITE (Turkish)

UMUX-LITE Items in Turkish [3]
Bu sistemin işlevleri gereksinimlerimi karşılamaktadır.
Bu sistemin kullanımı kolaydır.

Author contributions All Authors: Collaboratively conducted the comprehensive data analysis and prepared the result tables. All authors were involved in the study's conceptualization, provided critical revisions, and approved the final version of the manuscript. S.B.: Responsible for preparing Figs. 1, 2 and 3, conducted data analysis, contributed to manuscript writing, and participated in the review process. N.B.A.: Conducted the survey, contributed to data collection and initial data analysis, and participated in manuscript writing and review. B.S.: Conducted the survey, contributed to data collection and initial data analysis, and participated in manuscript writing and review.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Adam, M., Wessel, M., Benlian, A.: AI-based chatbots in customer service and their effects on user compliance. Electron. Markets. **31**(2), 427–445 (2021). https://doi.org/10.1007/s12525-020-00414-7
2. Akıl, S., Ungan, M.C.: E-Commerce logistics service quality. J. Electron. Commer. Organ. **20**(1), 1–19 (2021). https://doi.org/10.4018/jeco.292473
3. Berkman, M., S., Şahin, Ş.: Exploring usability as a formative construct through UMUX: A multi-language approach for Turkish adaptation. Int. J. Hum Comput Interact. 1–25 (2022). https://doi.org/10.1080/10447318.2022.2121049
4. Borsci, S., Federici, S., Bacci, S., Gnaldi, M., Bartolucci, F.: 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. Hum Comput Interact. **31**(8), 484–495 (2015). https://doi.org/10.1080/10447318.2015.1064648
5. Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., Chamberlain, A.: The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. Personal. Uniquit. Comput. **26**(1), 95–119 (2022). https://doi.org/10.1007/s00779-021-01582-9
6. Borsci, S., Prati, E., Malizia, A., Schmettow, M., Chamberlain, A., Federici, S.: Ciao AI: The Italian adaptation and validation of The chatbot usability scale. Personal. Uniquit. Comput. **27**(6), 2161–2170 (2023)
7. Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., Van Der Velde, F.: A confirmatory factorial analysis of the chatbot usability scale: A multilanguage validation. Personal. Uniquit. Comput. **27**(2), 317–330 (2023). https://doi.org/10.1007/s00779-022-01690-0
8. Brooke, J.: SUS: A quick and dirty usability scale. In: Usability Evaluation in industry, 1st edn., pp. 207–212. CRC (1996). https://doi.org/10.1201/9781498710411-35
9. Cole, D.A.: Methodological contributions to clinical research utility of confirmatory factor analysis in test validation research. J. Consult. Clin. Psychol. **55**(4), 584–594 (1987). https://doi.org/10.1037/0022-006X.55.4.584
10. Diederich, S., Brendel, A.B., Morana, S., Kolbe, L.: On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. J. Association Inform. Syst. **23**(1), 96–138 (2022). https://doi.org/10.17705/1jais.00724
11. Finstad, K.: The usability metric for user experience. Interact. Comput. **22**(5), 323–327 (2010). https://doi.org/10.1016/j.intcom.2010.04.004
12. Følstad, A., Brandtzæg, P.B.: Chatbots and the new world of HCI. Interactions. **24**(4), 38–42 (2017)
13. Guner, H., Acarturk, C.: The use and acceptance of ICT by senior citizens: A comparison of technology acceptance model (TAM) for elderly and young adults. Univ. Access Inf. Soc. **19**(2), 311–330 (2020). https://doi.org/10.1007/s10209-018-0642-4
14. Hair, J., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis, 7th edn. Pearson Prentice Hall, Upper Saddle River, New Jersey (2010)
15. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat. Lang. Eng. **6**(3–4), 287–303 (2000). https://doi.org/10.1017/s1351324900002497
16. Hu, L.T., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct. Equation Modeling: Multidisciplinary J. **6**(1), 1–55 (1999)
17. İçgözü, T.: Covid-19 etkisiyle chatbot kullanımı bankacılıkta 5 kat, e-ticarette 2 kat arttı. Webrazzi. (2020). Retrieved from https://webrazzi.com/2020/05/20/pandemi-etkisiyle-chatbot-kullanimi-bankacilikta-5-kat-e-ticarette-2-kat-artti/ Accessed May 18, 2023
18. Jenneboer, L., Herrando, C., Constantinides, E.: The impact of chatbots on customer loyalty: A systematic literature review. J. Theoretical Appl. Electron. Commer. Res. **17**(1), 212–229 (2022). https://doi.org/10.3390/jtaer17010011
19. Lewis, J.R.: Standardized questionnaires for voice interaction design. Voice Interact. Des. **1**(1), 1–16 (2016)
20. Lewis, J.R., Sauro, J.: Three Questionnaires for Measuring Voice Interaction Experiences. MeasuringU. (2020)
21. Lewis, J.R., Utesch, B.S., Maher, D.E.: UMUX-LITE: when there's no time for the SUS. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 2099–2102). (2013), April
22. Li, C.H.: The performance of ML, DWLS, and ULS Estimation with robust corrections in structural equation models with ordinal variables. Psychol. Methods. **21**(3), 369 (2016)
23. Liu, W., Yao, M.: Gender identity and influence in human-machine communication: A mixed-methods exploration. Comput. Hum. Behav. **144**, 107750 (2023)
24. Mariani, M.M., Hashemi, N., Wirtz, J.: Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. J. Bus. Res. **161**, 113838 (2023). https://doi.org/10.1016/j.jbusres.2023.113838
25. Mindrila, D.: Maximum likelihood (ML) and diagonally weighted least squares (DWLS) Estimation procedures: A comparison of Estimation bias with ordinal and multivariate non-normal data. Int. J. Digit. Soc. **1**(1), 60–66 (2010)
26. Polkosky, M.D.: Toward a social-cognitive psychology of speech technology: Affective responses to speech-based e-service. [University of South Florida]. In Digital Commons. (2005). http://ovidsp.ovid.com/ovidweb.cgi?T=JS%26PAGE=reference%26D=psyc4%26NEWS=N%26AN=2005-99018-095
27. Press, G.: AI Stats News: Chatbots Increase Sales By 67% But 87% Of Consumers Prefer Humans. Forbes. (2019). Retrieved from https://www.forbes.com/sites/gilpress/2019/11/25/ai-stats-news-chatbots-increase-sales-by-67-but-87-of-consumers-prefer-humans/?sh=6477b17d48a3 Accessed May 5, 2023
28. Ren, Q., Li, S., Song, B., Chen, C.: The application of bounded online gradient descent algorithms for kernel based online learning in tourist number forecasting. Proceedings of the International Conference on Electronic Business (ICEB), Guilin, China, 800–804. (2018), December
29. Salza, P.L., Foti, E., Nebbia, L., Oreglia, M.: MOS and pair comparison combined methods for quality evaluation of text-to-speech systems. Acta Acustica United Acustica. **82**(4), 650–656 (1996)
30. Seeger, A.-M., Heinzl, A.: Human versus machine: Contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In Lecture Notes in Information Systems and Organisation (Vol. 25, pp. 129–139). Springer Heidelberg. (2017). https://doi.org/10.1007/978-3-319-67431-5_15
31. Seeger, A.-M., Pfeiffer, J., Heinzl, A.: Texting with Human-like conversational agents: Designing for anthropomorphism. J. Association Inform. Syst. **22**(4) (2021). https://doi.org/10.17705/1jais.00685

32. Seo, S.: When female (Male) robot is talking to me: Effect of service robots' gender and anthropomorphism on customer satisfaction. Int. J. Hospitality Manage. **102** (2022). https://doi.org/10.1016/j.ijhm.2022.103166
33. Tavakol, M., Dennick, R.: Making sense of Cronbach's alpha. Int. J. Med. Educ. **2**, 53–55 (2011). https://doi.org/10.5116/ijme.4dfb.8dfd
34. Taylor, M.P., Girard, S., Jacobs, K., Buvat, J., Subrahmanyam, K., Puttur, R., Shah, H.: & B., A. Smart Talk: How Organizations and Consumers are Embracing Voice and Chat Assistants. In Capgemini. (2020). https://www.capgemini.com/wp-content/uploads/2019/09/Report_Conversational-Interfaces-1.pdf