

Chinese version of the chatbot usability scale: cross-cultural adaptation and validation

Haoming Ma, Runyuan Pei, Sijia Li, Aoqi Wang, Xingyi Tang, Meihua Piao

Submitted to: JMIR Human Factors
on: September 28, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4
Supplementary Files..... 24
 Multimedia Appendixes 25
 Multimedia Appendix 1..... 25



Chinese version of the chatbot usability scale: cross-cultural adaptation and validation

Haoming Ma¹ PhD; Runyuan Pei¹; Sijia Li¹; Aoqi Wang¹; Xingyi Tang¹; Meihua Piao¹ PhD

¹Beijing School of Nursing, Chinese Academy of Medical Sciences & Peking Union Medical College Beijing CN

Corresponding Author:

Meihua Piao PhD

Beijing

School of Nursing, Chinese Academy of Medical Sciences & Peking Union Medical College

No.33 Badachu Road, Shijingshan province

Beijing

CN

Abstract

Chatbots are increasingly deployed across domains, yet systematic evaluation of their usability remains limited, particularly in non-Western contexts. The 11-item Chatbot Usability Scale (BUS-11) has shown strong psychometric properties in prior studies, but no validated Chinese version exists despite China being one of the largest chatbot markets. This study aimed to translate, culturally adapt, and validate BUS-11 for Chinese users. Following established cross-cultural adaptation procedures, the scale was forward-and back-translated, reviewed by an expert committee, and pilot-tested for clarity and feasibility. A main validation study was then conducted with 146 participants who completed 438 chatbot evaluations across ten widely used systems. Psychometric analyses demonstrated excellent content validity (S-CVI = 0.92), strong internal consistency (Cronbach's α = 0.923), and a clear three-factor structure (Accessibility, Interaction Process Quality, Information Quality) explaining 56.1% of the variance, while Privacy/Security and Response Time were retained as single-item indicators. The Chinese BUS-11 proved concise (completion time <5 minutes), user-friendly, and psychometrically robust. This work fills a critical gap by providing the first validated instrument for assessing chatbot usability in Chinese contexts, enabling reliable cross-cultural comparisons and supporting both research and practical design evaluation in human-computer interaction.

(JMIR Preprints 28/09/2025:84971)

DOI: <https://doi.org/10.2196/preprints.84971>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in http://preprints.jmir.org/preprint/84971

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in https://doi.org/10.2196/preprints.84971

Original Manuscript

Chinese version of the chatbot usability scale: cross-cultural adaptation and validation

Haoming Ma¹, Runyuan Pei¹, Sijia Li¹, Aoqi Wang¹, Xingyi Tang¹, Meihua Piao^{1,*}

1 School of Nursing, Chinese Academy of Medical Sciences & Peking Union Medical College, No.

33 Ba Da Chu Road, Shijingshan District, Beijing, 100144, China

*Correspondence: Meihua Piao parkmihua@snu.ac.kr

Abstract

Chatbots are increasingly deployed across domains, yet systematic evaluation of their usability remains limited, particularly in non-Western contexts. The 11-item Chatbot Usability Scale (BUS-11) has demonstrated strong psychometric properties in prior studies; however, a validated Chinese version does not exist, despite China being one of the largest chatbot markets. This study aimed to translate, culturally adapt, and validate BUS-11 for Chinese users. Following established cross-cultural adaptation procedures, the scale was forward- and back-translated, reviewed by an expert committee, and pilot-tested for clarity and feasibility. A main validation study was then conducted with 146 participants who completed 438 chatbot evaluations across ten widely used systems. Psychometric analyses demonstrated excellent content validity (S-CVI = 0.92), strong internal consistency (Cronbach's $\alpha = 0.923$), and a clear three-factor structure (Accessibility, Interaction Process Quality, Information Quality) explaining 56.1% of the variance, while Privacy/Security and Response Time were retained as single-item indicators. The Chinese BUS-11 proved concise (completion time <5 minutes), user-friendly, and psychometrically robust. This work fills a critical gap by providing the first validated instrument for assessing chatbot usability in Chinese contexts, enabling reliable cross-cultural comparisons and supporting both research and practical design evaluation in human-computer interaction.

Keywords: Chatbot; Usability; Satisfaction; User experience; Cross-Cultural Adaptation



Introduction

Chatbots, also known as conversational agents, are artificial intelligence systems designed to interact with users through natural language dialogue(1). They have become increasingly ubiquitous across the globe in domains ranging from online retail and customer service to public administration and healthcare(2-4). For example, Alibaba's "AliMe" chatbot processed 300 million queries in one day during the 2019 Singles' Day shopping festival, government portals across 70% of China's provinces deployed chatbots by 2019, and hundreds of hospitals now employ chatbots for patient guidance(5-7). These examples illustrate the ubiquity of chatbots in China. Given this widespread adoption, ensuring chatbot usability is critical, as it directly impacts user satisfaction, acceptance, and sustained use(8).

In the context of general human-computer interaction(HCI), usability refers to how easy, efficient, and satisfying a system is for users to achieve their goals(9). For chatbots specifically, usability concerns the quality of the interaction experience, including coherence, appropriateness of responses, and perceived trustworthiness(10). High usability is crucial because it directly affects user satisfaction and continued use of chatbot services(11). However, traditional usability measures such as the System Usability Scale were designed for graphical user interfaces and often fail to capture conversational aspects unique to chatbots, such as turn-taking, naturalness of dialogue, and contextual coherence(12). Recognizing this gap, Borsci and colleagues developed the 11-item Chatbot Usability Scale (BUS-11), which demonstrated strong psychometric properties across several language (10, 13). Yet, no validated version of BUS-11 currently exists in Chinese, despite China being one of the largest chatbot markets.

Cross-cultural adaptation of usability scales is a recognized challenge in HCI. Established guidelines, such as Beaton et al. (2000) for instrument adaptation, emphasize the importance of semantic, conceptual, and cultural equivalence to ensure valid measurement across contexts(14). Prior efforts to localize usability instruments, such as Chinese versions of the System Usability Scale, highlight both feasibility and the need for culturally sensitive adaptation(15). Building on this foundation, our work focuses on adapting the BUS-11 for Chinese users.

Moreover, with the rapid emergence of AI-driven chatbots, including large language model-based systems such as ChatGPT and Deepseek, there is a growing demand in the HCI community for robust tools to evaluate chatbot usability and user satisfaction(16). The lack of validated instruments for The present study addresses this gap with three key contributions: (1) a rigorous translation and cultural adaptation of BUS-11 into Chinese, (2) empirical validation of its reliability, validity, and factor structure through a study involving 146 participants and 438 chatbot evaluations, and (3) new insights into Chinese users' perceptions of chatbot usability, laying the groundwork for cross-cultural comparisons in chatbot User Experience (UX) research.

Methods

1. Scale Translation and Cultural Adaptation

Following authorization from the original scale's corresponding author (Chamberlain et al.), the translation of the

Chatbot Usability Scale (BUS-11) into Chinese was conducted using the Brislin translation model as improved by Jones et al(17). Consistent with established guidance for cross-cultural instrument adaptation (e.g., forward-back translation with expert reconciliation), we documented all translation decisions to ensure semantic and conceptual equivalence. The specific steps are as follows: (a) Forward Translation: The researcher and a bilingual professional proficient in both Chinese and English independently translated the original BUS into Chinese. Discrepancies and alternative phrasings were recorded in a reconciliation log and resolved through consensus, resulting in an initial Chinese draft; (b) Back Translation: Another bilingual translator, whose native language is Chinese, translated the initial Chinese draft back into English, producing a back-translated English version; (c) Consistency Check: A monolingual English-speaking expert familiar with the BUS-11 reviewed and compared the original scale with the back-translated version. Any residual inconsistencies (e.g., colloquialisms or culture-specific expressions) were discussed in joint meetings with all translators until consensus was achieved, ensuring conceptual and semantic consistency; (d) Final Translation: A final Chinese translation was prepared collaboratively by all translators, incorporating adjustments from the consistency review, thus establishing semantic equivalence with the original BUS-11. Minor wording refinements were made for clarity and idiomatic naturalness (no item content was substantively altered).

Expert Committee Review (Cultural Adaptation)

In June 2024, six experts were invited to assess the cultural appropriateness of the translated scale. Each expert reviewed the Chinese translation against the original BUS-11, focusing on semantic equivalence, conceptual equivalence, idiomatic expressions, and readability. The panel comprised specialists in HCI/usability, informatics/AI, and scale development; feedback was synthesized and used to adjust item wording where needed (e.g., simplifying technical terms and aligning tone with typical Chinese conversational norms). The resulting version was designated the “final Chinese BUS-11” for subsequent pilot and main studies.

2. Pilot Study (Task Design, Feasibility, and Clarity Check)

Rationale and Procedure

To ensure participants evaluated chatbots after a realistic, standardized interaction, we designed representative use-case tasks for 10 commonly used chatbot systems in China. For example, participants were instructed to interact with Taobao's intelligent customer service, simulating a typical query related to membership benefits: “Your friend recommends you purchase Taobao's 88vip. Please use Taobao's chatbot 'Smart Customer Service Xiao Mi' to help you understand the specific benefits of 88vip. Log in as instructed and converse with this chatbot.” See the supplementary file for details. Immediately after completing the assigned task, participants filled out the Chinese BUS-11 about that chatbot. The same task-then-rating flow was later applied in the main study.

Sample and Data Collection

In July 2024, a pilot study involving 15 participants was conducted at a Beijing-based company, using convenience sampling. Participants completed the Chinese BUS-11 after performing designated tasks with selected chatbots. Completion times were recorded, and qualitative feedback regarding item clarity, ambiguity, and comprehension difficulties was collected through structured interviews. The pilot's primary aims were to assess task feasibility, confirm average completion time, and identify any items requiring wording refinements. Pilot data were used only for refinement and were not merged with the main validation dataset.

Outcomes Informing Quality Control

Based on the observed average completion time (~5 min 20 s), we set a priori a data-quality rule for the main study to exclude online questionnaires completed in <5 minutes, in addition to removing any incomplete submissions.

3. Content Validity Validation

Between June and July 2024, eight experts with backgrounds in human-computer interaction, informatics, and scale development evaluated the content validity of the Chinese BUS-11. Experts rated the relevance of each item to its intended measurement dimension on a four-point scale. Item-level content validity index (I-CVI) was computed as the proportion of experts assigning a relevance rating of 3 or 4 to an item; the scale-level content validity index (S-CVI) was calculated as the average of I-CVIs across items. Following common criteria, items with I-CVI ≥ 0.78 and overall S-CVI/Ave ≥ 0.90 were considered to exhibit acceptable to excellent content validity. Content validity was quantified using Item-level Content Validity Index (I-CVI) and Scale-level Content Validity Index (S-CVI), ensuring comprehensive item coverage and minimal redundancy. All content validity analyses were conducted before the main validation study.

4. Main Validation Study: Participants, Measures, Procedure, and Analysis

Participant Recruitment

Participants were recruited via social media platforms (e.g., Red Note, WeChat) according to predefined inclusion criteria: native Chinese speakers or individuals proficient in Chinese and aged 18 or above. Recruitment followed an online convenience-sampling approach; participation was voluntary. Each participant was randomly assigned to interact with three distinct chatbots (selected from the 10 systems), performing one predefined task per chatbot, and completed BUS-11 immediately after each interaction. Thus, each participant contributed three independent chatbot evaluations, yielding a repeated-measures data structure (evaluations nested within participants). For psychometric analyses (e.g., reliability and EFA), evaluations were analyzed at the evaluation level (total $n = 438$); design implications of repeated measures are discussed in the limitations.

Sample Size Estimation

Based on factor analysis requirements (5-10 times the number of items, accounting for a 20% invalid rate), a minimum sample size of 69 participants was established. We targeted ≥ 70 participants and exceeded this target, ultimately recruiting 146 participants who provided 438 evaluations, thereby improving the robustness of reliability and factor-analytic estimates.

Measures.

Two primary instruments were utilized:

- (i) General demographic questionnaire: collected participant demographics and chatbot usage patterns.
- (ii) Chinese BUS-11: a culturally adapted 11-item scale assessing chatbot usability across five dimensions. Because two BUS-11 dimensions ("Privacy and Security" and "Response Time") are single-item constructs in the original instrument, they were retained in the survey but treated as single-item indicators in factor-analytic planning (see Analysis).

Data Collection

Following ethical approval by the Ethics Committee at Peking Union Medical College School of Nursing, data were collected electronically via Tencent Questionnaire. Participants completed tasks on their own devices and then accessed the BUS-11 through a survey link; electronic informed consent was obtained before participation. Strict criteria ensured data quality: Questionnaires that were not fully completed were directly discarded. Based on pilot timing, online questionnaires completed in <5 minutes were excluded to reduce insufficient-effort responses. In the pilot experiment, the research team measured the time taken to complete the questionnaire, with an average duration of 5 minutes and 20 seconds. The finalized main-study dataset comprised 438 valid chatbot evaluations.

Data Analysis

Data were exported to Excel, verified, and analyzed using IBM SPSS Statistics 26.0. Reliability was evaluated using Cronbach's α coefficients for the overall scale and for each of its multi-item dimensions. Structural validity was assessed via exploratory factor analysis (EFA) after sampling-adequacy checks (Kaiser–Meyer–Olkin index and Bartlett's test of sphericity). Factors were extracted using principal components with the Kaiser criterion (eigenvalues > 1.0), guided by scree-plot inspection, and rotated using varimax for interpretability; loadings ≥ 0.40 were considered salient. In line with the instrument's structure, the EFA focused on the 9 items pertaining to multi-item dimensions; the two single-item dimensions ("Privacy and Security," "Response Time") were not included in EFA but were retained as single-item indicators in descriptive analyses. P-values < .05 were considered statistically significant for inferential tests (e.g., Bartlett's). We note that PCA is commonly used in early-stage exploratory work for data reduction and structure discovery; future work may complement this with common-factor methods (e.g., principal axis factoring) and confirmatory factor analysis once larger samples are available.

Results

1. Participant Characteristics

We analyzed 438 valid chatbot evaluations submitted by 146 participants (69 male, 145 female), with each participant randomly assigned to evaluate three of ten popular Chinese chatbot systems using the Chinese BUS-11. The sample skewed young (67.3% aged 18–24; 22.4% aged 25–30), was highly educated (92.9% held a bachelor's degree or above), and represented multiple industries, most prominently healthcare (52.8%) and education/training (11.7%). Most respondents had prior exposure to chatbots (79.0% have used a chatbot) and at least moderate familiarity (43.0% reported being familiar/very familiar), while 27.5% reported being unfamiliar/very unfamiliar. These characteristics suggest a relatively tech-savvy cohort, which should be considered when interpreting generalizability. Detailed distributions are provided in Table 1.

Table 1. Characteristics of participants

	Item	Frequency	Proportion (%)
Gender	Male	69	32.2
	Female	145	67.8
Age (years)	18~24	144	67.3
	25~30	48	22.4
	31~40	14	6.5
	41~50	5	2.3
	51~60	3	1.4
	Junior High School and Below	2	0.9
Educational Level	High School	13	6.0
	Bachelor's Degree	121	56.5

Industry	Master's Degree	78	36.4
	and Above		
	Manufacturing	9	4.2
	Construction	4	1.9
	Education/Training	25	11.7
	Internet	5	2.3
	Healthcare	113	52.8
	Computer	11	5.1
	Finance	6	2.8
	Scientific Research	9	4.2
Familiarity with Chatbots	Government	10	4.7
	Institutions		
	Professional		
	Consulting	6	2.8
	Services		
	Culture and		
	Entertainment	8	3.7
	Others	8	3.7
	Very Unfamiliar	26	12.1
	Unfamiliar	33	15.4
Have you used a Chatbot before?	Neutral	63	29.4
	Familiar	61	28.5
	Very Familiar	31	14.5
	No	45	21.0
	Yes	169	79.0
Frequency of Using Chatbots Weekly	Never (=0)	53	24.8
	Rarely (1-2)	69	32.2
	Occasionally (3-4)	46	21.5
	Often (5 - 6)	25	11.7
	Every day (=7)	21	9.8

2. Scale Reliability and Validity

Content Validity

Eight domain experts (HCI, informatics, chatbot engineering, and scale development) rated item relevance on a 4-point scale. Background information of the experts is presented in Table 2. Item-level content validity indices (I-CVIs) ranged from 0.875 to 1.000; four items (Items 3, 4, 6, 11) achieved I-CVI = 1.000 (unanimous relevance). The scale-level CVI (S-CVI/Ave) was 0.920, exceeding common acceptability thresholds (≈ 0.80), indicating excellent content validity of the Chinese BUS-11.

Table 2. Background information of the experts

Expert	Gender	Age	Education	Years of Experience	Position	Title
--------	--------	-----	-----------	---------------------	----------	-------

1	Male	37	Doctorate	5	Doctoral Supervisor	Researcher
2	Male	38	Doctorate	3	Master's Sup ervisor	Associate Researcher
3	Male	31	Doctorate	1	Master's Supervisor	Associate Researcher
4	Male	33	Doctorate	6	Master's Supervisor	Associate Researcher
5	Male	50	Doctorate	27	Doctoral Supervisor	Professor
6	Male	36	Doctorate	12	Master's Supervisor	Associate Researcher
7	Female	42	Doctorate	15	Doctoral Supervisor	Associate Researcher
8	Male	34	Doctorate	8	Systems Engineer	Senior Engineer

Structural Validity

Before factor extraction, data suitability was confirmed (KMO = 0.743; Bartlett's χ^2 (36) = 628.185, $p < .001$). Because two BUS-11 dimensions are single-item constructs (Privacy/Security; Response Time), they were not included in the EFA; the analysis therefore used the remaining 9 items (see Table 3; scree plot in Figure 1).

Table 3 data suitability for EFA

Test Item	Value
KMO Measure of Sampling Adequacy	0.743
Bartlett's Test of Sphericity	Approximate Chi-Square Degrees of Freedom Significance
	628.185 36 □0.001

In this study, principal component analysis and varimax orthogonal rotation were used to extract common factors. 9 items (items 10 and 11 were not included in the confirmatory factor analysis due to having only one item each) were analyzed, and combined with the scree plot (Figure 2), 3 common factors with eigenvalues greater than 1.0 were ultimately extracted, with a cumulative variance contribution rate of 56.107 %. The results of the exploratory factor analysis for this experiment are shown in Table 3. 2. 3.

Figure 2. Scree plot of the factors

Exploratory factor analysis with varimax rotation revealed a clear three-factor structure for the Chinese version of BUS-11 (Table 3). Factor 1 (Accessibility) consisted of two items reflecting the ease of locating and using chatbot functions, with strong loadings (.840 and .760). Factor 2 (Interaction Process Quality) captured three items describing conversational smoothness, coherence, and comprehensibility, with loadings ranging from .716 to .759. Factor 3 (Information Quality) included four items related to understanding user intent, appropriateness of information, absence of irrelevant content, and reliability of responses, with loadings between .627 and .739.

Table 3 Factor loadings of the Chinese version of BUS-11

Dimension	Item	Factor				
		1	2	3	4	5
Access the chatbot's functions	Finding the entry to this chatbot is an easy task	0.840				
	It is easy to find the various functions of this chatbot. (e.g., uploading file analysis or changing chat avatar, etc.)	0.760				
Quality of chatbot functions	The communication process with this chatbot is smooth		0.757			
	This chatbot is able to maintain context coherence		0.759			
	The responses from this chatbot are easy to understand		0.716			
Provide quality of conversation and information	I feel this chatbot understands my intentions, and assists me in achieving the purpose of this chat			0.683		
	The information provided by this chatbot is appropriate, neither too much nor too little			0.627		
	This chatbot does not provide information irrelevant to my needs			0.629		
	I find the responses of this chatbot to be accurate and reliable			0.739		
Privacy and Security	This chatbot clearly informs me of the privacy policy and potential risks				1	
Response Time	With a good internet connection, the response time of the chatbot is short					1

Internal Consistency (Reliability)

The overall internal consistency of the Chinese BUS-11 was excellent (Cronbach's $\alpha = 0.923$). By factor: Accessibility $\alpha = 0.788$; Interaction Process Quality $\alpha = 0.910$; Information Quality $\alpha = 0.931$. Reliability is not computed for the single-item dimensions (Privacy/Security; Response Time).

Discussion

Rigorous Localization and Cultural Adaptation

This study undertook a rigorous localization of the 11-item ChatBot Usability Scale (BUS-11) for the Chinese context, providing a valuable tool for evaluating chatbot usability in China. Following established best practices for cross-cultural instrument adaptation, we employed a forward-backward translation procedure inspired by the Brislin model with iterative expert committee review(18). Each step—from forward translation to back-translation and reconciliation—was conducted systematically to ensure linguistic accuracy and conceptual equivalence. The expert panel, comprising specialists in HCI, human factors, and related fields, scrutinized the translation for semantic consistency, idiomatic appropriateness, and cultural relevance. This thorough process yielded a Chinese version of BUS-11 that is semantically and conceptually aligned with the original, with no major changes to item content needed. The fact that no substantial modifications were required suggests that BUS-11's core constructs (e.g., perceived *Accessibility*, *Information Quality*, etc.) are robust and transferable across cultures, reflecting a degree of universality in chatbot usability dimensions. Moreover, by carefully documenting the translation decisions and cultural adjustments, our work enhances transparency and reproducibility for future researchers. This level of methodological detail not only provides *traceability* for our localization (which future studies can build upon or audit) but also offers practical insights for others adapting user experience scales across languages. Given the rapid adoption of chatbot systems in diverse domains worldwide(19), the introduction of a standardized Chinese-language usability instrument is a timely contribution. It enables researchers and practitioners in China to rigorously assess chatbot user experience and compare findings with international studies, thereby integrating Chinese user data into the broader HCI knowledge base. In summary, the localization process was conducted in a scientifically rigorous manner, ensuring the Chinese BUS-11 achieves a high degree of semantic and cultural equivalence to the original. This approach exemplifies the emphasis on careful user-centric methodology in human-computer interaction research and aligns with human-centered design principles that call for understanding users in context(20).

Participant Sample and Representativeness

Our sample selection aimed for diversity in gender, age, education, and industry background, resulting in a broad cross-section of Chinese chatbot users. In practice, however, the achieved sample was skewed toward younger, highly educated adults: most participants were 18–30 years old, and over 90% held at least a bachelor's degree. While this profile likely reflects the early adopters of new technologies(21) and those most engaged with contemporary chatbot applications, it also introduces some limitations on generalizability. Younger and tech-savvy

users may have different usability perceptions than older or less technologically experienced individuals. Prior research in human factors has noted that older adults can face unique usability challenges and may interact with technology in distinct ways(22). Our current sample underrepresents such groups, as well as users from rural areas or lower educational backgrounds, which means the usability impressions captured by the scale might not fully reflect the broader population of potential chatbot users. On the positive side, the concentration of experienced users in our sample could heighten sensitivity to subtle differences in chatbot interaction quality—these users are likely familiar with a variety of interfaces and may provide discerning feedback. The inclusion of participants from a range of industries (education, healthcare, internet/IT, etc.) is another strength, as it suggests the scale items were interpreted meaningfully across different usage contexts. Indeed, the BUS-11 was designed to be domain-agnostic, and our results tentatively support its applicability in multiple sectors. However, some fields (e.g. finance, construction) were represented by relatively few respondents, so conclusions about those domains should be drawn cautiously. Likewise, because recruitment relied on convenience and snowball sampling (primarily in a few urban regions), there may be regional or cultural sub-group differences within China that our study did not capture. In sum, our sample provides a solid initial test of the scale among active chatbot users, but it is not fully representative of all user segments. Future research should broaden the sampling frame to include older adults, less frequent technology users, and a wider geographic distribution to ensure the Chinese BUS-11 is robustly validated for the general population. Broadening the participant base in this way would enhance the scale’s universality and address the “Western, educated, industrialized, rich, democratic” (WEIRD) bias analogue in the Chinese context, thus strengthening confidence that the tool works well for *all* target users, not just the young and educated subset.

Reliability, Validity, and Key Psychometric Findings

We conducted a comprehensive psychometric evaluation of the Chinese BUS-11, including assessments of content validity, construct (factorial) validity, internal consistency reliability, and user feedback on item clarity and practical usability. Overall, the findings provide strong evidence that the localized scale retains the scientific integrity of the original instrument. Here we discuss each set of findings in turn, relating them to established benchmarks and prior research, and interpret their implications for the scale’s use in HCI and human factors studies.

Content Validity

By consulting a panel of eight experts to review the relevance of each item, we established that the content validity of the Chinese BUS-11 is high. The item-level Content Validity Index (I-CVI) for every item was very strong, with four items (Items 3, 4, 6, and 11) achieving an I-CVI of 1.00 (i.e., unanimously deemed highly relevant by all experts). The remaining items also scored well above conventional acceptability thresholds(23). Given that an I-CVI ≥ 0.78 is often considered the minimum when six or more experts are involved, our results indicate that each item is considered clearly representative of the intended construct by domain specialists. The scale-level content validity index (S-CVI) was 0.920 (both by the averaging method and the universal agreement method), exceeding the common cutoff of 0.90 for excellent content coverage of a construct. These indices confirm that the BUS-11 (Chinese version) has *comprehensive* content coverage: the items collectively cover the important facets of chatbot usability without obvious omissions. This outcome was facilitated by the diverse expertise of our review panel,

which included professionals in human-computer interaction, usability engineering, and systems engineering. Their varied perspectives helped ensure that the items were not only translated correctly but also conceptually appropriate for Chinese users. In practical terms, the high content validity suggests that Chinese BUS-11 can confidently be used to gauge chatbot usability aspects as intended—such as efficiency, clarity of answers, and user comfort—without missing key elements. The rigorous content validation process, in line with recommended scale development procedures, provides a solid foundation for the subsequent construct validation. It also offers reassurance to practitioners that the instrument has face validity in the local context (the items *make sense* to experts and presumably to end-users as well). We believe this strong content validity will help drive adoption of the scale in both research and industry evaluations, as stakeholders can trust that the instrument measures what it is supposed to measure (i.e., salient user experience factors for chatbots).

Structural Validity (Exploratory Factor Analysis)

To examine the underlying factor structure of the Chinese BUS-11, we carried out an exploratory factor analysis (EFA) on the survey responses. Preliminary checks confirmed that our data were suitable for factor analysis: the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was 0.743, which is above the usual threshold of 0.60 and can be considered “middling” to “good”(24), and Bartlett’s test of sphericity was highly significant, indicating sufficient inter-item correlations for extracting latent factors. Using principal component extraction with varimax rotation, we identified a three-factor solution that best fit the data. These three extracted common factors had eigenvalues greater than 1 and together accounted for about 56.1% of the total variance in users’ responses. A cumulative variance above ~50% is acceptable in behavioral research, given the complexity of human-computer interaction measures, so 56% indicates that the scale captures a substantial portion of the usability perception variance in our sample. Each factor showed a clear thematic grouping of items, reflecting distinct dimensions of the chatbot user experience. Notably, two specific usability aspects—Privacy/Security and Response Time—were each originally measured by a single item in BUS-11. As single-item facets, they could not load onto multi-item factors in the EFA and were thus analyzed separately (we examined their scores independently rather than including them in the factor structure). The remaining 9 items clustered into three coherent factors: (i) Accessibility (e.g., ease of accessing and using the chatbot, such as simple interface and low effort to start the interaction), (ii) Interaction Quality (covering the smoothness, coherence, and understandability of the conversational exchange), and (iii) Information Quality (assessing the relevance, clarity, and correctness of the chatbot’s responses, without unnecessary redundancy). This structure is broadly consistent with the original conceptual framework of the BUS scale, which encompassed multiple dimensions of chatbot usability(10), but our findings suggest a somewhat more consolidated factor model in the Chinese context.

In the original development of BUS, a five-factor model (BUS-15) was proposed, which included separate dimensions for privacy and response speed alongside factors analogous to those we found(10). Our EFA results imply that Chinese users may not distinguish as many separate categories, instead perceiving a more integrated set of usability factors. Attributes related to the conversational process and functional outcome appear to intertwine; users who find the interaction process smooth also tend to perceive the information provided as high quality,

suggesting an underlying general perception of “interaction effectiveness.” Interestingly, a recent re-examination of BUS-11 in a large multi-chatbot dataset also identified a simpler factor structure than initially theorized(25). Our findings align with that perspective, reinforcing the idea that a parsimonious model (fewer broader factors) can sufficiently capture chatbot usability evaluations. It is worth noting that we chose to retain the Privacy/Security and Response Time items as standalone indicators in the instrument, due to their importance in usability (users do care about privacy and speed) and their presence in the original scale. However, from a psychometric standpoint, single-item factors are suboptimal because they do not allow for internal consistency reliability estimation and can’t capture the breadth of a construct. Best practices in scale development typically recommend having at least 2–3 items per factor to achieve adequate reliability and construct representation(26). Therefore, one implication is that future refinement of the Chinese BUS might involve expanding these facets—adding a couple of items to better gauge the privacy/security aspect of chatbot use (e.g., covering data protection, user consent, etc.) and the responsiveness aspect (e.g., not just speed but also perceived *responsiveness* or promptness). For now, including those single items ensures that our localized scale does not lose any content relative to BUS-11, and they provide useful standalone measures (for instance, designers might specifically want to see ratings of privacy transparency). In summary, the EFA supports the structural validity of the Chinese BUS-11 by confirming that its items coherently measure multiple distinct dimensions of usability. The identified three-factor structure captures the major themes of chatbot usability experience in our sample. This structure provides an empirical basis for scoring or interpreting the scale (e.g. computing subscale scores for Accessibility, Interaction Quality, and Information Quality) and sets the stage for future confirmatory testing to verify whether this structure generalizes to other samples or holds under more stringent statistical criteria.

Internal Consistency and Reliability

The Chinese BUS-11 demonstrated high internal consistency, indicating that the items within each identified dimension reliably measure the same underlying concept. The overall Cronbach’s alpha for the 11-item scale was $\alpha = 0.923$, which is well above the conventional threshold of 0.70 for acceptable reliability in early-stage research(27). This suggests that the instrument produces stable and consistent scores, making it suitable for both research and practical evaluation purposes. Looking at the subscales corresponding to our three main factors: Accessibility items yielded a subscale alpha of approximately 0.79, Interaction Quality items around 0.91, and Information Quality items around 0.93. These figures range from good to excellent in terms of reliability (with $\alpha > 0.90$ indicating excellent internal consistency in many social science contexts). The relatively high alphas for Interaction Quality and Information Quality reflect that users’ ratings on those grouped items moved together closely – in other words, participants who reported that the chatbot was easy to converse with also tended to rate its answers as clear and appropriate, and vice versa. This pattern is consistent with the idea that these facets are related in practice (as discussed above in the factor analysis interpretation). Meanwhile, the Accessibility subscale’s alpha being above 0.7 is acceptable and suggests a moderate consistency among those items (which typically cover aspects like interface clarity and ease of initiating the chat). It is worth noting that we did not perform a test–retest reliability check (i.e., measuring the same users at two different times) due to the cross-sectional nature of our data collection. As a result, we cannot directly comment on the temporal stability of the scores. Future work should

assess test–retest reliability to ensure the scale yields consistent results over time when the underlying usability of a system has not changed. Nonetheless, the combination of a coherent factor structure and strong internal consistency in this study provides convergent evidence of the scale’s reliability. In practical terms, someone using this Chinese BUS-11 can be reasonably confident that the items are reliable indicators of each usability dimension – if a chatbot improves (or worsens) in one aspect according to the scale, all items related to that aspect will likely reflect the change. These reliability findings are comparable to those reported in the original BUS scale development(10) and other standardized usability questionnaires, lending further credibility to the adapted scale’s psychometric soundness.

User Feedback on Item Clarity and Scale Usability

Beyond quantitative metrics, we gathered qualitative feedback from participants regarding the questionnaire’s clarity and ease of use. Overall, users reported that the Chinese BUS-11 was easy to understand and that its items covered the core aspects of their chatbot usage experience. The average time to complete the survey was around 3–4 minutes, indicating that the scale imposes a minimal burden on respondents. This brevity is an important advantage for usability instrumentation: similar standard questionnaires like the System Usability Scale (10 items) typically take only a few minutes to fill out(28), and even ultrashort measures like UMUX-LITE (with just 2 items) are valued for quick administration (29). Our 11-item scale strikes a balance between comprehensiveness and efficiency. A completion time under five minutes means it can be deployed in iterative design cycles or field studies without causing respondent fatigue or distraction, thereby encouraging higher response rates and more thoughtful answers. Practitioners could, for example, embed the BUS-11 in a post-chatbot-interaction feedback form or as part of a usability test, confident that users can complete it very quickly. This ease of use aligns with the general HCI goal of minimizing *cognitive load* on users wherever possible – here, the survey itself respects that principle by not overtaxing the participant(30).

Importantly, while feedback on item clarity was mostly positive, we did identify one area of concern: the item regarding “Privacy policy and potential risk disclosure” was interpreted inconsistently by some users. A few participants commented that they were unsure what exactly was being asked, or that they hadn’t considered the issue until seeing the question. This variability in understanding likely reflects differences in users’ familiarity with or attention to privacy and security issues. Prior research has shown that many users tend not to thoroughly read or recall online privacy policies and terms of service(31), which can lead to low awareness of what “privacy risk disclosure” entails in practice. Additionally, individuals vary in privacy *literacy* and concern – for instance, those with higher privacy awareness often show greater concern about data practices(32). In our sample, some users may have been very cognizant of chatbot privacy implications (and thus welcomed the item), whereas others paid it little heed, resulting in different interpretations of the question. This finding suggests that the phrasing of the privacy/security item might need refinement to ensure all users interpret it consistently. We plan to address this by possibly providing a brief explanation or example within the item (e.g., “...such as the chatbot clearly explaining how your data will be used”) to anchor respondents’ understanding. Clarifying the wording can reduce the cognitive effort required to comprehend the question and make the intended meaning more transparent(30). In turn,

this should improve the reliability of responses on that item. On a broader level, this issue highlights that user perceptions of privacy are complex and may not be as immediately tangible as more straightforward usability aspects like speed or accuracy. It reinforces the idea that continuous user feedback and iterative tweaking of survey language are crucial – especially when dealing with abstract concepts like trust, risk, and privacy in technology use. Despite that one semantic issue, users generally felt that the BUS-11 was comprehensive yet concise. Several participants noted that the questions were relevant to their interaction experience. The content validity exercise with experts appears to have paid off from the end-user perspective as well: none of the participants suggested that an obvious usability aspect was missing from the survey. This gives us confidence that the localized scale not only *appears* valid to experts on paper but also resonates with actual users' experiences. Taken together, the user feedback underscores the BUS-11's practicality: it is easy to fill out, takes only a few minutes, and produces data that users themselves find reflective of their experience. Such user-friendliness is essential for deployment at scale. For instance, companies deploying chatbots can use this questionnaire to gather systematic UX feedback after a chat session without inconveniencing customers. Likewise, research studies can include the BUS-11 alongside other measures (e.g., task success rates or qualitative interviews) without overburdening participants. The positive reception in terms of clarity also means that we can be more confident in our quantitative results – when respondents understand the questions well, their answers are more trustworthy. In summary, the BUS-11 (Chinese) proved to be a user-friendly instrument with mostly clear items, though ongoing refinement (especially for the privacy item) will be important to further enhance clarity and ensure all facets of chatbot usability are intuitively understood by respondents of varying backgrounds.

Limitations

While this study yielded encouraging results for the Chinese adaptation of BUS-11, several limitations must be acknowledged to contextualize the findings and guide future work. First, our sample, as discussed, was not fully representative of all user demographics. The geographic distribution of participants was relatively narrow (with many users from a few major cities), and certain age groups (especially adults over 50) and occupations were underrepresented. This means the current validation is most directly applicable to younger, educated Chinese users in urban settings. Caution should be used in generalizing the psychometric results to, say, older adults or rural users, who might interact with chatbots differently or have different usability concerns. Expanding the sample in future studies will be important to verify that the scale maintains its reliability and validity across a broader swath of the Chinese population. Additionally, larger and more heterogeneous samples would allow for analyses of measurement invariance—checking whether the scale functions equivalently across subgroups such as older vs. younger users, or high-frequency vs. low-frequency chatbot users(33).

Second, certain BUS-11 dimensions in our current form rely on a single item (notably the Privacy/Security and Response Time facets). Single-item measures, while convenient, provide limited psychometric information. They can neither capture the full breadth of a construct nor estimate internal consistency for that construct. The strong performance of the overall scale notwithstanding, the precision and reliability of those facets could be improved. Future research should consider enriching these dimensions by developing and testing additional items. For

example, the privacy facet might be expanded with items addressing whether the chatbot clearly asks permission for data usage or whether it provides settings for privacy control. The response time facet might include an item on whether the chatbot's speed meets user expectations or if delays ever cause frustration. Of course, any new items would themselves need to undergo translation and validation. The trade-off is that adding items could slightly increase completion time, but as long as the total remains reasonable (under, say, 2 minutes increase), it would likely be worth it for a more robust scale. In short, our current findings highlight the need to bolster those one-item constructs for a more rigorous measurement model(26).

Third, our study primarily evaluated the scale in general usage scenarios (common chatbot systems for generic tasks). We did not deeply examine how the scale performs in specific industry contexts or use-case scenarios. It remains an open question whether users in specialized domains — such as healthcare, education, finance, or customer service — interpret and prioritize the BUS-11 items similarly. Different contexts may put extra emphasis on certain usability aspects; for instance, in healthcare chatbot applications, issues like empathy, trust, and accuracy of information might be even more critical(34), whereas in e-commerce chatbots, speed and transactional clarity could dominate user satisfaction. Our inability to test such variations is a limitation. It's possible that the BUS-11 might require minor tweaks or weighting adjustments to be optimally effective in certain domains, or it may prove to be robust as-is — we simply don't have the data yet to say. Future validations should include a variety of chatbot types and domains to ensure the scale's content validity and structure hold in those settings, or to highlight any domain-specific gaps that could be addressed by additional context-specific items.

Finally, our psychometric analysis relied on exploratory factor analysis due to the novelty of applying BUS-11 in Chinese. We have not yet conducted a confirmatory factor analysis (CFA) on a separate sample, which would be the next step to confirm whether the three-factor model (plus the two single-item factors) holds under stricter modeling conditions. A CFA would allow us to test the hypothesized factor structure (possibly including correlated factors or a higher-order factor of overall usability) and to adjust for measurement errors(35). It would also enable formal comparison of alternative models – for example, whether a structure with five distinct factors (if privacy and speed were expanded) fits significantly better than a three-factor structure, or whether a single-factor model (treating usability as unidimensional) is decisively worse than a multi-factor model. Additionally, with a large enough sample, multi-group CFA could be used to test measurement invariance across key groups (e.g., male vs female, younger vs older, different regions), to ensure the scale operates equivalently(36). Because our current sample size was modest for CFA and was all collected as one group, we leave these confirmatory analyses for future work. In summary, our use of EFA was appropriate for an initial validation, but further confirmation via CFA is needed to solidify the factor structure and to verify that the Chinese BUS-11 meets the standards of psychometric validity expected for a widely applicable research instrument.

Conclusion

This study delivers a validated Chinese adaptation of BUS-11 with excellent internal consistency, strong expert-rated content validity, and a clear three-factor structure for multi-item facets (Accessibility, Interaction Process Quality, Information Quality), while preserving the privacy and speed facets conceptually. The instrument fills a

measurement gap for chatbot UX evaluation in China and provides researchers and practitioners with a concise, psychometrically sound tool suitable for laboratory and field contexts alike. Future work should establish temporal stability, convergent validity, and confirmatory structure, and expand coverage to underrepresented user groups and domain-specific scenarios to further strengthen its utility and generalizability.

Abbreviations

CFA: Confirmatory factor analysis

HCI: Human–computer interaction

I-CVI: Item-level content validity index

LLM: Large language models

S-CVI: Scale-level content validity index

UX: User experience

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the School of Nursing, Peking Union Medical College (approval number: PUMCSON-2023-10). All the participants provided informed consent.

Consent for publication

The authors consent to publication.

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Disclosure statement

The authors declare no conflicts of interest.

Funding

This study was funded by the Non-Profit Central Research Institute Fund of the Chinese Academy of Medical Sciences (Grant NO. 2023-RC320-01).

Reference

1. Nagarhalli TP, Vaze V, Rana N, editors. A review of current trends in the development of chatbot systems. 2020 6th International conference on advanced computing and communication systems (ICACCS); 2020: IEEE.
2. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health*. 2019;5:2055207619871808.
3. Androutsopoulou A, Karacapilidis N, Loukis E, Charalabidis Y. Transforming the communication between citizens and government through AI-guided chatbots. *Government information quarterly*. 2019;36(2):358-67.
4. Landim A, Pereira A, Vieira T, de B. Costa E, Moura J, Wanick V, et al. Chatbot design approaches for

fashion E-commerce: an interdisciplinary review. *International Journal of Fashion Design, Technology and Education*. 2022;15(2):200-10.

5. Li F-L, Qiu M, Chen H, Wang X, Gao X, Huang J, et al., editors. *Alime assist: An intelligent assistant for creating an innovative e-commerce experience*. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*; 2017.

6. Wang Y, Zhang N, Zhao X. Understanding the determinants in the different government AI adoption stages: Evidence of local government chatbots in China. *Social Science Computer Review*. 2022;40(2):534-54.

7. Shan Y, Ji M, Xie W, Zhang X, Qian X, Li R, et al. Use of health care chatbots among young people in china during the omicron wave of COVID-19: evaluation of the user experience of and satisfaction with the technology. *JMIR Human Factors*. 2022;9(2):e36831.

8. Silva GRS, Canedo ED. Towards user-centric guidelines for chatbot conversational design. *International Journal of Human-Computer Interaction*. 2024;40(2):98-120.

9. Hornbæk K. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*. 2006;64(2):79-102.

10. Borsci S, Malizia A, Schmettow M, Van Der Velde F, Tariverdiyeva G, Balaji D, et al. The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and ubiquitous computing*. 2022;26(1):95-119.

11. Ren R, Zapata M, Castro JW, Dieste O, Acuña ST. Experimentation for chatbot usability evaluation: A secondary study. *IEEE Access*. 2022;10:12430-64.

12. Ren R, Castro JW, Acuña ST, De Lara J. Evaluation techniques for chatbot usability: A systematic mapping study. *International Journal of Software Engineering and Knowledge Engineering*. 2019;29(11n12):1673-702.

13. 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 and Ubiquitous Computing*. 2023;27(6):2161-70.

14. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25(24):3186-91.

15. Wang Y, Lei T, Liu X. Chinese system usability scale: Translation, revision, psychological measurement. *International Journal of Human-Computer Interaction*. 2020;36(10):953-63.

16. Bolpagni M, Gabrielli S, editors. Development of a comprehensive evaluation scale for LLM-powered counseling chatbots (CES-LCC) using the Edelpi method. *Informatics*; 2025: MDPI.

17. Jones PS, Lee JW, Phillips LR, Zhang XE, Jaceldo KB. An adaptation of Brislin's translation model for cross-cultural research. *Nursing research*. 2001;50(5):300-4.

18. Brislin RW. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*. 1970;1(3):185-216.

19. Brandtzaeg PB, Følstad A. Chatbots: changing user needs and motivations. *interactions*. 2018;25(5):38-43.

20. Norman D. *The design of everyday things: Revised and expanded edition*: Basic books; 2013.

21. Rogers EM, Singhal A, Quinlan MM. *Diffusion of innovations. An integrated approach to communication theory and research*: Routledge; 2014. p. 432-48.

22. Charness N, Boot WR. Aging and information technology use: Potential and barriers. *Current directions in psychological science*. 2009;18(5):253-8.

23. Lynn MR. Determination and quantification of content validity. *Nursing research*. 1986;35(6):382-6.

24. Kaiser HF. An index of factorial simplicity. *psychometrika*. 1974;39(1):31-6.

25. Borsci S, Schmettow M. Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots. *Personal and Ubiquitous Computing*. 2024;28(6):1033-44.

26. Worthington RL, Whittaker TA. Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist*. 2006;34(6):806-38.

27. Nunnally JC. An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook*. 1978:97-146.

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

29. Lewis CC, Fretwell CE, Ryan J, Parham JB. Faculty use of established and emerging technologies in

higher education: A unified theory of acceptance and use of technology perspective. *International Journal of Higher Education*. 2013;2(2):22-34.

30. Sweller J. Cognitive load during problem solving: Effects on learning. *Cognitive science*. 1988;12(2):257-85.
31. McDonald AM, Cranor LF. The cost of reading privacy policies. *Isjlp*. 2008;4:543.
32. Baruh L, Secinti E, Cemalcilar Z. Online privacy concerns and privacy management: A meta-analytical review. *Journal of Communication*. 2017;67(1):26-53.
33. Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*. 2000;3(1):4-70.
34. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*. 2018;25(9):1248-58.
35. Brown T. *Confirmatory factor analysis for applied research*, 2nd ed. 2015.
36. Vandenberg, Robert J, Lance, Charles E. *A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research*. *Organizational Research Methods*. 2000.

Supplementary Files

Multimedia Appendixes

Standardized user tasks applied across ten popular chatbot systems in China to ensure consistency and ecological validity in usability evaluations with the Chinese BUS-11.

URL: <http://asset.jmir.pub/assets/0d6e2b3c5f1284216981ae6f06beb2ff.docx>