An Indonesian Adaptation of the Bot Usability Scale Full Version (BUS-42)

Harry Budi Santoso*
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
harrybs@cs.ui.ac.id

Budiarti
Ministry of Education, Culture, Research, and Technology
Central Jakarta, Indonesia
budiarti@kemdikbud.go.id

Martin Schrepp
SAP SE
Baden-W ̈urttemberg, Germany
martin.schrepp@sap.com

Mira Suryani
Faculty of Mathematics and Natural Sciences
Universitas Padjadjaran,
West Java, Indonesia
mira.suryani@unpad.ac.id

Rizal Fathoni Aji
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
rizal@cs.ui.ac.id

Erza Janitradevi Nadine
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
erza.janitradevi@ui.ac.id

Dony Abdul Chalid
Faculty of Economic and Business
Universitas Indonesia
Depok, Indonesia
donyabdul@ui.ac.id

Abstract

The Bot Usability Scale (BUS-42) is a tool specifically designed to evaluate chatbot usability, yet it has not been extensively applied in Indonesian contexts. This study aimed to adapt the BUS for Indonesian users, following Beaton's cross-cultural adaptation framework to ensure cultural and linguistic relevance. The Indonesian adaptation was tested with 103 participants from the Faculty of Computer Science at one of large public universities in Indonesia, resulting in a high Cronbach's Alpha of 0.922, confirming strong internal consistency. Theoretical implications of the findings is an adapted version of the BUS-42 that is culturally relevant toward Indonesian, but still has the same meaning as the original version. Therefore, the practical implications are that the adapted BUS-42 can be used to evaluate chatbot usability that are commonly used in Indonesian to identify areas for improvement of the chatbots towards Indonesian user needs. Future research should focus on extending adapted versions of BUS-42 testing to diverse chatbot applications and optimize its utility across various Indonesian user contexts.

Keywords: Bot Usability Scale, BUS, chatbot usability, cross-cultural adaptation, Indonesian adaptation, reliability, questionnaire translation

* Corresponding Author

Introduction

AI-powered chatbots have become essential tools that enhance user interactions across various digital platforms. Chatbots are designed to mimic human conversation by providing relentless support and services across customer service, education, and healthcare, highlighting their significance in the contemporary digital ecosystem (Tsakiridis et al., 2020). As chatbots become more prevalent globally, evaluating their usability in diverse cultural and linguistic settings becomes increasingly critical. A systematic review by Borsci et al. (2022) that analyzed instruments used to evaluate the usability of chatbots showed that the Bot Usability Scale (BUS) is more effective in evaluating usability compared to the Usability Metrics for User Experience in terms of the completeness of its questionnaire questions.

The advantage of BUS is its focus on measuring the usability of AI-based chatbots by focusing on elements of user interaction with the chatbot, such as naturalness of dialogue, ease of following conversation flow, perceived usefulness, and trust in the bot. This aims to identify user pain points when using chatbots, thereby improving their usability. A study by Borsci et al. (2023), which adapted BUS to the Italian version shows that the adaptation facilitates easier usability evaluation for chatbots because it has been adapted for context and language. This process is known as cross-cultural adaptation. Through cross-cultural adaptation, respondents still grasp the original meaning of the questionnaire, but in a more accessible language version.

According to a study by Hadiyati & Tovtora S (2025) which shows that the use of AI-based chatbots in Indonesia is experiencing rapid growth, shows that there is a notable lack of locally adapted instruments to evaluate chatbot usability. This presents an opportunity for researchers to adapt the BUS questionnaire to an Indonesian version. Therefore, this research addresses this gap by adapting the original BUS version, BUS-42, to align with Indonesia's linguistic and cultural context. Adaptation of the original BUS version with 42 instruments also aims to ensure that the scale captures t relevant dimensions of chatbot usability. By focusing on the complete scale, the study aims to provide a robust foundation for understanding and improving chatbot usability in Indonesia (Noviyanti et al., 2021) . By localizing the BUS-42 for Indonesian audiences, this study contributes to the broader discourse on enhancing AI's interaction with users across different cultural backgrounds and languages. It aims to equip Indonesian practitioners and researchers with a robust tool to assess and optimize chatbot interactions, leading to enhanced user satisfaction and engagement in digital environments.

This paper is organized into five sections. The Introduction presents the background and objectives of the study. The Literature Review section discusses key concepts essential for understanding the research, including usability, usability evaluation methods, the Bot Usability Scale, approaches to cross-cultural adaptation, and fundamental aspects of questionnaire quality. The Adaptation of BUS-42 section details the adaptation process and the development of the Indonesian version of BUS-42. Finally, the Conclusions and Further Work section summarizes the key findings, research limitations, and future directions.

Literature Review

Usability

Usability refers to the ease and efficiency with which individuals can achieve their objectives using a product or system, while also ensuring a positive experience. It involves assessing how effectively users can accomplish their tasks with a service, typically evaluated through structured research methods known as "usability testing," which measures factors such as success rates and user satisfaction. While UX encompasses the overall design and experience of a product, usability specifically focuses on optimizing functionality to ensure the product works seamlessly for the user (Digital.gov, 2025). It is also recognized as a quality attribute that evaluates how easy user interfaces are to use (Nielsen, 1994).

Usability consists of five key components: 1) learnability, which measures how easily users can complete tasks during their first encounter with the system; 2) efficiency, which assesses how quickly users can accomplish tasks after learning the system; 3) memorability, which evaluates how easily users can retain their ability to use the system after a period of not using it; 4) errors, which considers the

number of mistakes users make, the severity of those mistakes, and how easily users can recover from them; and 5) satisfaction, which gauges how pleasant the system is for users to use (Nielsen, 1994). In addition to these quality components, usability is also divided into three principles: learnability, flexibility, and robustness (Dix et al., 2003). Learnability refers to how easily new users can understand the system's interface, flexibility relates to the variety of ways users can interact with the system, and robustness is concerned with how well the system supports users in solving problems that arise during use.

Usability Evaluation

Usability evaluation refers to the methods and processes employed to assess the usability of a product or system. It typically involves user testing, where participants are observed interacting with the system to identify usability issues and gather data on user performance and satisfaction. Methods for usability evaluation range from formal techniques such as usability testing, heuristic evaluation, and cognitive walkthroughs, to more informal approaches like surveys and interviews (Rubin & Chisnell, 2008). The goal is to uncover any obstacles to efficient and satisfactory user interaction and to provide actionable insights for improving the product.

In the context of chatbots, usability evaluation is crucial for understanding how users interact with conversational interfaces and for ensuring that these systems meet the diverse needs of users across different cultures and languages (Bennett et al., 2025; Kuric et al., 2025). These methods allow researchers to gather actionable data of chatbot users on task completion rates, error frequency, and user satisfaction, which are essential for refining the product (Kuric et al., 2025).

On the other hand, expert-based evaluations, such as heuristic evaluations and cognitive walkthroughs, involve usability experts reviewing the system based on established usability principles. These evaluations help predict potential user problems early in the development process, making them valuable for identifying significant usability issues before user testing begins (Nielsen, 1994). Remote usability testing also has gained popularity, allowing researchers to gather data from users in their natural environments, which is especially useful for products intended for a global audience. Automated tools also contribute to usability evaluation by tracking user behavior, and providing insights through heatmaps and session recordings that complement traditional methods (Tullis & Albert, 2013).

A secondary study by (Ren et al., 2022) analyzed previous research to identify experimental methods used to evaluate chatbot usage to gain a detailed understanding of their effectiveness, efficiency, and user satisfaction. This study aimed to understand the dimensions that need to be considered when evaluating chatbot usability. (Hidayat et al., 2022) discussed in their study that in the educational context, usability evaluation of chatbots is also useful for determining whether chatbots can be used in education to support the learning process. Therefore, this usability evaluation can be used to assess the extent to which chatbots can provide an effective and accessible learning experience, while also identifying areas that still need improvement so that chatbots can truly function as optimal learning media.

Bot Usability Scale (BUS)

The Bot Usability Scale (BUS) is a specialized instrument designed to evaluate the usability of chatbots, addressing the unique challenges and interaction patterns inherent to these digital agents. As chatbots have become increasingly prevalent across various sectors—including customer service, healthcare, education, and e-commerce—the need for a standardized and reliable method to assess their usability has become crucial. The BUS was developed in response to this demand, offering a systematic framework to evaluate key aspects of chatbot usability, such as effectiveness, efficiency, and user satisfaction (Borsci et al., 2022). The Bot Usability Scale (BUS) was developed to address the lack of standardized tools for evaluating chatbot usability, particularly the unique conversational aspects that differentiate chatbots from other digital systems. Inspired by the System Usability Scale (SUS), the BUS underwent a comprehensive development process, starting with a systematic review to identify 38 relevant attributes, which were refined through expert consultations and focus groups. The initial BUS-42, encompassing 42 items, was psychometrically tested with 480 participants interacting with various

chatbots. This iterative process, including Bayesian Exploratory Factor Analysis, reduced the scale to the 15-item BUS-15, ensuring strong reliability (alpha = .87) while preserving its core dimensions of usability, such as ease of use, conversational quality, and user privacy. The BUS not only serves as an evaluative tool but also as a guide for developers to enhance chatbot design, providing a robust framework for understanding and improving user interactions.

Drawing inspiration from the widely-used System Usability Scale (Brooke, 1996), the BUS is tailored specifically to the nuances of chatbot interaction, making it an essential tool for developers and researchers aiming to optimize chatbot performance. The scale consists of a series of statements that users rate based on their experience with the chatbot, covering dimensions like ease of use, the naturalness of interaction, and the chatbot's ability to understand and respond appropriately to user inputs. The statements can be rated by participants of a study with a five-point Likert scale with answer options of 1 ('Strongly Disagree') and 5 ('Strongly Agree'). These ratings provide a quantitative measure of the chatbot's usability, highlighting areas that may require improvement. The BUS-42 addresses several critical dimensions of chatbot usability. The original BUS-42 is shown below in Table 1.

Table 1. Original BUS-42 developed by Borsci et al. (2022)

Attributes	**Items order**
Ease to start a conversation	It was clear how to start a conversation with the chatbot.
	It was easy for me to understand how to start the interaction with the chatbot.
	I find it easy to start a conversation with the chatbot.
	The chatbot was easy to access.
	The chatbot function was easily detectable.
	It was easy to find the chatbot.
Expectation setting	Communicating with the chatbot was clear.
	I was immediately made aware of what information the chatbot can give me.
	It is clear to me early on about what the chatbot can do.
Flexibility and communication effort	I had to rephrase my input multiple times for the chatbot to be able to help me.
	I had to pay special attention regarding my phrasing when communicating with the chatbot.
	It was easy to tell the chatbot what I would like it to do.
Ability to maintain a themed discussion	The interaction with the chatbot felt like an ongoing conversation.
	The chatbot was able to keep track of context.
	The chatbot maintained a relevant conversation.
Reference to the service	The chatbot guided me to the relevant service.
	The chatbot is using hyperlinks to guide me to my goal.
	The chatbot was able to make references to the website or service when appropriate.
Users' privacy and security	The interaction with the chatbot felt secure in terms of privacy.
	I believe the chatbot informs me of any possible privacy issues.
	I believe that this chatbot maintains my privacy.
Recognition and facilitation of users' goal and intent	I felt that my intentions were understood by the chatbot.
	The chatbot was able to guide me to my goal.
	I find that the chatbot understands what I want and helps me achieve my goal.
Relevance of information	The chatbot gave relevant information during the whole conversation.
	The chatbot is good at providing me with a helpful response at any point of the process.
	The chatbot provided relevant information as and when I needed it.

Attributes	Items order
Maxim of quantity	The amount of received information was neither too much nor too less.
	The chatbot gives me the appropriate amount of information.
	The chatbot only gives me the information I need.
Resilience to failure	The chatbot could handle situations in which the line of conversation was not clear.
	The chatbot explained gracefully when it could not help me.
	When the chatbot encountered a problem, it responded appropriately.
Understandability and politeness	I found the chatbot's responses clear.
	The chatbot only states understandable answers.
	The chatbot's responses were easy to understand.
Perceived conversational credibility	I feel like the chatbot's responses were accurate.
	I believe that the chatbot only states reliable information.
	It appeared that the chatbot provided accurate and reliable information.
Speed of answer	The time of the response was reasonable.
	My waiting time for a response from the chatbot was short.
	The chatbot is quick to respond.

The BUS has been applied in a previous study by Poglitsch et al. (2025), to evaluate and improve the usability of AI-based chatbot. Usability evaluation helped identify the chatbot's usability, demonstrating its effectiveness and engaging use in training social communication skills. However, there is still room for improvement, particularly in long-term context retention and tailoring emotional expressions to be more contextual. This demonstrates the potential for using BUS to evaluate chatbot usability in other contexts, such as education, industry, and enterprise.

Cross-Cultural Adaptation

Cross-cultural adaptation is a critical process in ensuring that assessment tools, such as the BUS, are valid and reliable across different cultural and linguistic contexts. The goal of cross-cultural adaptation is not merely to translate a tool from one language to another but to modify it in a way that preserves its conceptual meaning while making it culturally relevant and comprehensible to the target population. This process involves several stages, including translation, back-translation, expert committee review, pre-testing, and psychometric validation (Beaton et al., 2000). Each stage is essential for maintaining the tool's validity and reliability, ensuring that the adapted version accurately reflects the constructs it is intended to measure within the new cultural context.

In Indonesia, several usability and evaluation tools have undergone cross-cultural adaptation to meet local needs. For instance, tools like the System Usability Scale (SUS) (Sharfina & Santoso, 2016), E-learning Usability Scale (Hasibuan et al., 2020). Chatbot Usability Questionnaire (CUQ) (Noviyanti et al., 2021), User Experience Questionnaire (UEQ) (Mochammad Aldi Kushendriawan et al., 2021) and its extended version UEQ+ (Santoso et al., 2022), have been successfully adapted. These adaptations often involve addressing linguistic and cultural nuances, such as accommodating local idiomatic expressions, adapting examples to align with familiar technologies, and ensuring relevance to Indonesian users' technological literacy levels and communication styles. These adapted tools have enabled usability evaluations that are both meaningful and contextually appropriate for Indonesian users, further advancing the field of human-computer interaction in the region.

The first step in cross-cultural adaptation is translation, where the original tool is translated into the target language. This translation must go beyond literal word-for-word conversion, focusing on conveying the intended meaning in a way that is natural and clear in the target language. Following this, back-translation is conducted, where the translated version is retranslated into the original language by a different set of translators who are blind to the original version. This step is crucial for identifying

discrepancies between the original and translated versions and ensuring that the translation process has not altered the tool's meaning (Hambleton, 2005).

An expert committee then reviews the original, translated, and back-translated versions. This committee typically includes bilingual experts with knowledge of the tool's subject matter, as well as cultural and linguistic experts familiar with the target population. The committee's role is to resolve any inconsistencies, ensure cultural relevance, and refine the translation to better fit the target context. For instance, certain terms or phrases may need to be modified to align with cultural norms or everyday language used by the target population, ensuring that the tool is both understandable and relevant (Beaton et al., 2000).

Once the expert committee has finalized the adapted version, it undergoes pre-testing with a sample from the target population. Pre-testing involves administering the tool to a small group of individuals from the target culture to identify any remaining issues related to comprehension, interpretation, or cultural relevance. Feedback from pre-testing is used to make further adjustments, ensuring that the final version of the tool is clear and culturally appropriate. This step is particularly important for identifying any subtle cultural nuances that may have been overlooked during the translation and review process.

After pre-testing, the adapted tool undergoes psychometric validation to confirm its reliability and validity within the new cultural context. This involves statistical analysis to assess the tool's internal consistency, test-retest reliability, and construct validity. Ensuring that the adapted tool maintains its psychometric properties is critical for its effective use in research and practice. If the tool does not perform reliably or validly in the new context, it cannot be considered an effective measure of the constructs it was designed to assess (Hambleton, 2005).

In the context of chatbot usability, a cross-cultural adaptation of the BUS is essential for ensuring that the scale is applicable and meaningful across different cultural settings. Chatbots are increasingly used in global markets, where users come from diverse linguistic and cultural backgrounds. An adapted version of the BUS that accounts for these cultural differences can provide more accurate and relevant insights into how different user groups interact with and perceive chatbots. For example, cultural differences in communication styles, levels of technological literacy, and expectations of user interfaces can all influence how users experience chatbot interactions. By adapting the BUS to reflect these differences, researchers and developers can better assess and improve chatbot usability for a global audience.

Moreover, cross-cultural adaptation contributes to the broader goal of creating inclusive digital tools that cater to the needs of diverse user groups. In Indonesia, for instance, where there are multiple languages, dialects, and cultural practices, adapting the BUS to the local context ensures that it captures the unique aspects of Indonesian users' interactions with chatbots. This adaptation process not only enhances the validity of the usability assessments conducted with the BUS but also helps to generalize findings across different populations, allowing for a more comprehensive and equitable evaluation of chatbot systems worldwide.

Reliability Test

The reliability test focuses on determining whether a measurement instrument consistently and stably captures the phenomenon it is designed to measure. This is crucial for ensuring that the instrument exhibits high internal consistency, meaning that if the measurement is repeated under the same conditions, it should yield the same results (Taherdoost, 2016). Achieving reliability is fundamental to the trustworthiness of a research instrument, as it validates that the responses are not subject to random error (Huuck, 2013).

There are several methods available to measure reliability of an instrument. In UX questionnaires nearly all authors use Cronbach's Alpha (Cronbach, 1951) for this purpose. This is a simple metric based on the correlations of the items in the scale. The Alpha coefficient ranges theoretically from $-\infty$ to 1 (negative correlations can theoretically occur), but in practice the value will be between 0 (bad reliability) and 1 (high reliability). Higher values indicate greater reliability, with a value closer to 1

representing optimal internal consistency (George & Mallery, 2003) This study, calculated Cronbach's Alpha to verify that the adapted BUS-42 demonstrated high reliability, confirming its suitability for usability assessments in the Indonesian academic context.

Reliability is an essential characteristic of a research instrument, as it reflects the instrument's ability to consistently measure a concept or construct across repeated trials. For an instrument to be considered valid, it must first demonstrate reliability, as consistency is a prerequisite for meaningful measurement (Nunnally, 1978). Reliability testing assesses the stability and consistency of a survey or other measurement tools by calculating a reliability coefficient, which indicates how strongly the items within the instrument correlate with one another (Heffner et al., 2022).

One widely used statistic for assessing internal consistency is Cronbach's Alpha, which provides a single estimate of the instrument's reliability. Developed to measure the consistency of items within a test, Cronbach's Alpha evaluates whether all items contribute to measuring the same construct (Tavakol & Dennick, 2011). A value of 0.7 or above is generally accepted as an indicator of reliability, signifying that the items in the instrument are sufficiently correlated to measure a cohesive construct (George & Mallery, 2003). But please note that already Nunnally (1978) points out that the required level of Alpha depends on how critical the decisions are that are derived from the result of a questionnaire or test. Thus, there is no clear methodological foundation for any clear cut-points. Such suggestions are just conventions (Cortina, 1993; Schrepp, 2020).

Research Methodology

This study utilized a cross-cultural adaptation design based on the guidelines proposed by (Beaton et al., 2000) to adapt the BUS for Indonesian users. This adaptation involved three main phases: the Adaptation Phase, the Validation Phase, and the Reliability Testing Phase, each consisting of several systematic steps to ensure linguistic, conceptual, and cultural equivalency of the translated scale, as shown in Figure 1.

Study Design

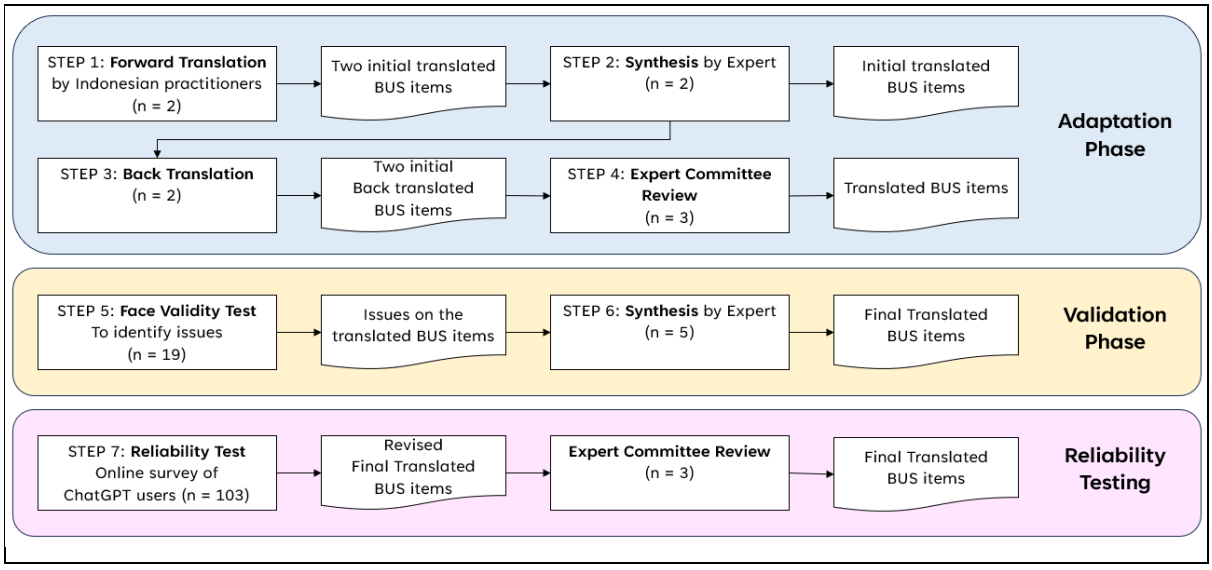

Figure 1. Cross-Cultural Adaptation Process of the Bot Usability Scale Full Version (BUS-42) for Indonesian Users

The Adaptation Phase began with Step 1: Forward Translation, where two Indonesian practitioners independently translated the original BUS-42 items into Indonesian. This step aimed to capture various linguistic nuances to ensure clarity and cultural relevance. In Step 2: Synthesis by Experts, two experts synthesized the two initial translations into a single version, producing the initial translated BUS-42 items. Step 3: Back Translation involved two bilingual translators who translated the synthesized items

back into the original language (English) to check for accuracy and consistency. Finally, Step 4: Expert Committee Review engaged a panel of three experts to review both the original and back-translated versions. This review ensured conceptual equivalency, linguistic accuracy, and cultural relevance in the translated items.

The Validation Phase aimed to evaluate the face validity of the translated BUS-42. In Step 5: Face Validity Test, an online survey was conducted with 19 participants from the target population to identify any issues related to item clarity and cultural relevance. Step 6: Synthesis by Experts followed, in which a group of five experts analyzed the feedback from the face validity test and refined the translated BUS-42 items, producing the final translated version.

The final phase, Reliability Testing, focused on assessing the reliability of the adapted BUS. In Step 7: Reliability Test, the final translated BUS-42 was distributed to a sample of 103 ChatGPT users in Indonesia via an online survey. This step provided data on the scale's internal consistency and reliability within the Indonesian context. Afterward, an Expert Committee Review with three experts was conducted to assess the overall reliability findings and ensure the final version of the BUS-42 was suitable for Indonesian users. This structured adaptation process, incorporating forward and back translations, expert synthesis, face validity testing, and reliability assessment, follows established guidelines to produce a culturally relevant and reliable usability scale for chatbot interactions.

Participants and Context of The Study

The participants in this study were students from the Faculty of Computer Science at one of large public universities in Indonesia selected due to their familiarity with digital technologies and chatbot interactions, which are relevant to the usability assessment. A total of 122 participants were involved in the adaptation and validation process, divided across two primary phases.

For the face validity test, 19 Master's students at the faculty participated. This phase was conducted offline using a questionnaire format to allow for direct interaction and detailed feedback. These participants provided essential insights into the clarity, cultural relevance, and comprehensibility of the translated BUS-42 items. Their feedback played a crucial role in identifying any linguistic or conceptual issues, which were then addressed to refine the scale, ensuring that it resonated well within the Indonesian cultural context.

In the reliability test phase, a number of 105 students participated. However, only 103 responses were valid and used for analysis. The participants were Undergraduate, Master's, and Doctoral levels in Computer Science at the faculty. This phase was conducted using an online questionnaire administered through Microsoft Forms, which allowed for efficient data collection from a diverse range of students. By including participants from different educational backgrounds, the study ensured a broader perspective on the usability scale, allowing for the assessment of the internal consistency and reliability of the final translated BUS-42 items across varying levels of experience and understanding within the field.

By selecting participants who were familiar with technology interfaces and experienced in digital communication, we aimed to ensure that the adapted Indonesian version of the BUS-42 would be relevant and applicable in real-world scenarios. Conducting research in this setting allowed us to rigorously test the cultural and linguistic appropriateness of the scale, as well as its reliability, ensuring that the final adapted BUS-42 would be a reliable and valid tool for evaluating chatbot usability in Indonesia. This approach strengthened the scale's applicability, making it a practical resource for usability assessment in Indonesia's evolving digital landscape.

Data Collection Procedures

The primary instrument used in this study was the Indonesian-adapted BUS-42, designed to evaluate chatbot usability across various dimensions. The BUS-42 covers a range of attributes, such as ease of conversation initiation, accessibility, user expectations, flexibility, ability to maintain themed discussions, service reference, privacy and security, goal recognition, information clarity, information sufficiency, failure resilience, comprehension, conversational credibility, and response speed. Each

attribute includes specific items that collectively measure the usability experience in chatbot interactions, making the BUS-42 a comprehensive tool for usability evaluation.

For this study, ChatGPT, a generative AI chatbot, was selected as a case study due to its widespread use and relevance in academic and practical applications. ChatGPT's extensive user base within the student population at the Faculty of Computer Science, Universitas X, one of large public universities in Indonesia, provided a robust context to assess the BUS-42's applicability. ChatGPT's functionalities, including answering questions, providing information, and supporting academic work, align with the usability dimensions the BUS-42 aims to measure.

Students were invited to participate in the study voluntarily. They were also asked to complete an informed consent form as an agreement to their participation. A number of 103 responses were valid and used for analysis. The demographic composition of the participants included Undergraduate, Master, and Doctoral students, providing diverse perspectives in the evaluation process. Among these participants, 65.05% were male and 34.95% were female. Usage patterns of ChatGPT among respondents varied, with 46.60% using it daily, 47.57% several times a week, and smaller percentages using it less frequently. Participants reported using ChatGPT for a range of purposes, including learning (51.46%), information search (28.16%), and academic assignments (12.62%).

Results and Discussion

The study aimed to validate the Indonesian-adapted BUS-42 for assessing chatbot usability, using ChatGPT as a case study. This adaptation followed Beaton's cross-cultural adaptation guidelines to ensure the scale's cultural and linguistic suitability for Indonesian users (Beaton et al., 2000). The Indonesian version of the BUS-42, shown in Table 2, comprises 42 items divided into key usability attributes, covering a wide range of chatbot usability dimensions such as ease of use, accessibility, flexibility, privacy, and response speed. These items allow for comprehensive measurement across various user experiences.

Table 2. The Indonesian Version of BUS-42

Attributes	**Items order**
Kemudahan memulai percakapan	Jelas bagi saya bagaimana cara memulai percakapan dengan *chatbot* ini.
	Mudah bagi saya untuk memahami bagaimana memulai interaksi dengan *chatbot* ini.
	Saya merasa mudah memulai percakapan dengan *chatbot* ini.
Akses ke Chatbot	*Chatbot* ini mudah diakses.
	Fitur-fitur *chatbot* mudah dideteksi.
	Saya mudah untuk menemukan *chatbot* ini.
Ekspektasi	Berkomunikasi dengan *chatbot* ini jelas.
	Saya langsung mengetahui informasi apa saja yang dapat diberikan *chatbot* untuk saya.
	Sejak awal, saya memahami apa yang bisa dilakukan oleh *chatbot* ini.
Fleksibilitas	Saya harus mengubah pertanyaan saya beberapa kali supaya *chatbot* ini dapat membantu saya.
	Saya harus memperhatikan pemilihan kalimat saya secara sungguh-sungguh saat berkomunikasi dengan *chatbot* ini.
	Mudah untuk memberitahu *chatbot* apa yang ingin saya lakukan.
Kemampuan untuk mempertahankan diskusi bertema	Interaksi dengan chatbot terasa seperti percakapan yang berkelanjutan.
	Chatbot mampu melacak konteks percakapan
	Chatbot mampu mempertahankan percakapan yang relevan.
Referensi ke layanan	*Chatbot* memandu saya menuju layanan yang relevan.
	Chatbot menggunakan tautan (link) untuk memandu saya menuju tujuan saya.
	Chatbot mampu memberikan rujukan ke situs atau layanan web apabila diperlukan.

Attributes	Items order
Privasi dan keamanan pengguna	Interaksi dengan *chatbot* terasa aman dari aspek privasi.
	Saya percaya *chatbot* memberi tahu saya tentang masalah privasi yang mungkin terjadi.
	Saya percaya *chatbot* menjaga privasi saya.
Pengenalan dan pemenuhan tujuan pengguna	Saya merasa bahwa *chatbot* memahami maksud saya.
	Chatbot mampu memfasilitasi saya untuk mencapai tujuan saya.
	Saya tahu bahwa *chatbot* memahami apa yang saya inginkan dan membantu saya untuk mencapai tujuan saya
Kejelasan Informasi	*Chatbot* memberikan informasi yang relevan sepanjang percakapan.
	Chatbot memberikan jawaban yang baik dalam membantu di setiap tahapan proses
	Chatbot memberikan informasi yang relevan pada saat saya membutuhkannya.
Kecukupan kuantitas	Jumlah informasi yang diterima tidak terlalu banyak atau terlalu sedikit.
	Chatbot memberikan saya jumlah informasi yang sesuai.
	Chatbot hanya memberikan informasi yang saya butuhkan.
Ketahanan terhadap kegagalan	*Chatbot* dapat menangani situasi ketika percakapan yang berlangsung kurang jelas.
	Chatbot menjelaskan dengan baik ketika tidak dapat membantu saya.
	Ketika chatbot menemui masalah, chatbot menjawab dengan tepat.
Pemahaman	Saya merasa jawaban *chatbot* jelas.
	Chatbot hanya memberikan jawaban yang mudah dipahami.
	Jawaban *chatbot* mudah dipahami.
Kredibilitas percakapan	Saya merasa jawaban dari *chatbot* akurat.
	Saya yakin bahwa *chatbot* hanya memberikan informasi yang dapat dipercaya.
	Dapat terlihat bahwa *chatbot* memberikan informasi yang akurat dan dapat diandalkan.
Kecepatan merespons	Waktu respons *chatbot* dapat ditoleransi.
	Waktu tunggu saya terhadap respons *chatbot* cukup singkat.
	Chatbot merespons dengan cepat.

The reliability of the Indonesian version of the BUS was measured using Cronbach's Alpha, which indicated excellent internal consistency with a score of 0.922 for 42 items, as shown in Table 3. This high Cronbach's Alpha score confirms that the scale is cohesive and effectively measures the intended usability construct, surpassing the standard threshold of 0.7 (George & Mallery, 2003; Tavakol & Dennick, 2011). The high Cronbach's Alpha score of 0.922 indicates strong internal consistency across the items in the BUS, affirming that the scale is reliable and effectively captures the usability construct for chatbots in the Indonesian context (Tavakol & Dennick, 2011). The comprehensive structure of the adapted BUS-42, as shown in Table 4, supports its use in academic settings by covering a wide range of usability attributes.

Table 3. The Cronbach's Alpha Result of the Indonesian Version of BUS

Cronbach's Alpha	Number of Items
0.922	42

Table 4. Item-Total Statistics of Indonesian Version of BUS

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach Alpha if Item Deleted
A1	0.458	0.27	0.93	0.88
A2	0.458	0.29	0.92	0.9
A3	0.455	0.31	0.93	0.86
B1	0.44	0.34	0.8	0.59
B2	0.46	0.24	0.86	0.68
B3	0.438	0.38	0.77	0.62
C1	0.38	0.62	0.76	0.63
C2	0.396	0.47	0.88	0.4
C3	0.394	0.6	0.74	0.75
D1	0.383	0.39	0.73	-0.39
D2	0.377	0.33	0.78	-0.49
D3	0.387	0.76	0.27	0.64
E1	0.406	0.44	0.77	0.78
E2	0.405	0.44	0.81	0.62
E3	0.411	0.41	0.85	0.52
F1	0.284	0.92	0.72	0.78
F2	0.32	0.73	0.85	0.56
F3	0.313	0.68	0.85	0.58
G1	0.296	0.94	0.88	0.75
G2	0.306	0.10	0.83	0.84
G3	0.307	0.95	0.89	0.72
H1	0.409	0.3	0.82	0.71
H2	0.398	0.32	0.79	0.77
H3	0.41	0.27	0.88	0.6
I1	0.394	0.42	0.89	0.77
I2	0.396	0.43	0.87	0.82
I3	0.392	0.45	0.88	0.78
J1	0.357	0.57	0.82	0.62
J2	0.349	0.65	0.77	0.66
J3	0.365	0.51	0.83	0.65
K1	0.359	0.71	0.87	0.76
K2	0.347	0.79	0.83	0.81
K3	0.36	0.72	0.89	0.7
L1	0.367	0.54	0.61	0.7
L2	0.391	0.32	0.82	0.57
L3	0.369	0.38	0.85	0.26

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach Alpha if Item Deleted
M1	0.296	0.75	0.78	0.81
M2	0.321	0.58	0.86	0.74
M3	0.302	0.62	0.89	0.61
N1	0.424	0.48	0.87	0.88
N2	0.425	0.36	0.94	0.81
N3	0.425	0.41	0.91	0.83

Insights from usage patterns showed that most participants used ChatGPT frequently for academic-related purposes, aligning with usability attributes such as information clarity and flexibility, which are essential for educational applications. Consistent usability scores across demographic groups further suggest that the BUS provides stable, reliable measurements across varied profiles, indicating its suitability for diverse educational contexts.

The adaptation process was conducted to ensure conceptual equivalence and linguistic naturalness, not just a literal translation. For example, for the attribute 'Resilience to failure', the item 'explained gracefully' was adapted to 'explain well'. A literal translation such as 'explain gracefully' would be highly unnatural and distort the meaning. This adaptation ensured the item measured the correct concept in a way that was easily understood by Indonesian respondents, thus improving the quality of the instrument. The findings from this case study suggest that the Indonesian-adapted BUS is a reliable tool for assessing chatbot usability in contexts where generative AI is used extensively, such as educational and research settings. These findings validate the Indonesian-adapted BUS -42 as a robust tool for assessing chatbot usability, with practical implications for evaluating AI-driven tools like ChatGPT in academic and potentially other Indonesian settings. The study underscores the importance of culturally adapting usability scales to ensure meaningful assessments in diverse user environments (Beaton et al., 2000).

Conclusion and Future Work

This study successfully created and validated the Indonesian-adapted BUS-42 as a reliable tool for assessing chatbot usability within an academic setting. By following Beaton’s cross-cultural adaptation guidelines, the scale was made culturally relevant and linguistically precise for Indonesian users. Reliability analysis yielded a high Cronbach’s Alpha score of 0.922, indicating excellent internal consistency across 42 items. This finding demonstrates the scale's robustness and coherence in measuring chatbot usability, establishing the adapted BUS-42 as a valuable tool for evaluating user experience on AI-driven platforms like ChatGPT in Indonesia.

The study's findings confirmed that the adapted BUS-42 captures a comprehensive range of usability attributes, such as ease of use, accessibility, information clarity, privacy, and response speed. The high usage frequency of ChatGPT among participants, primarily for educational and research purposes, highlights the growing relevance of chatbots in academic environments. The scale’s consistent performance across different demographic groups, including gender and educational level, further underscores its versatility and potential for broader applications.

The theoretical implications of this research include providing an adapted version of the BUS that is culturally relevant, precise, and has the same meaning as the original version. Meanwhile, the practical implications are that this adapted BUS questionnaire can be used to evaluate chatbot usability in local contexts more accurately and efficiently, support designers and developers in identifying areas for improvement, and provide a practical and standardized instrument for research and implementation in various fields such as education, health, public services, and digital business. As the use of chatbot innovation continues to grow in Indonesia, future studies may build upon the findings of this research by employing an adapted version of established chatbot evaluation instruments. Such an approach

would enable the implementation of contextually, culturally, and linguistically appropriate assessments of chatbot usability and effectiveness across both educational and industrial settings.

As limitations we need to mention that we used students as participants to check if the translated items of the BUS-42 are easy to understand and for the final validation study. Of course, our two student samples are not representative for all persons that might interact with chatbots. Our participants are of course younger than the average Indonesian inhabitant and have a higher level of education. Thus, further checks and potentially improvements of the items with more representative samples might improve the wording of the items further.

Although the BUS-42 has shown significant potential as a usability assessment tool, its use has so far been limited in scope. For the BUS-42 to reach its full potential, it should be adopted and tested across various other types of chatbots, such as those in e-commerce, government services, and messaging platforms. Expanding its application could reveal the scale's adaptability and effectiveness in diverse functionalities and user contexts.

Further linguistic testing of the Indonesian-adapted BUS-42 could enhance its clarity for a broader demographic, ensuring the accessibility of the language and improving the precision of usability feedback. Additionally, to optimize its applicability, future research could incorporate validity and sensitivity tests, which would help verify the scale's accuracy in capturing nuanced usability feedback and detecting subtle differences in user experience across different chatbot interactions. These additional validations and expansions would strengthen the BUS-42 as a reliable tool for usability evaluation, making it valuable for a wider range of user-centered chatbot development efforts in Indonesia.

Acknowledgements

This research was supported by Direktorat Riset dan Pengembangan (Risbang) Universitas Indonesia through Hibah Publikasi Terindeks Internasional (PUTI) Q1 2024—2025 (Grant Number: NKB-133/UN2.RST/HKP.05.00/2024).

References

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine*, *25*(24), 3186–3191. https://doi.org/10.1097/00007632-200012150-00014

Bennett, R. J., Tsiolkas, J., & Tagudin, J. (2025). Usability and desirability of a hearing health chatbot: an explorative study. *International Journal of Audiology*, 1–11. https://doi.org/10.1080/14992027.2025.2514586

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Borsci, S., Prati, E., Malizia, A., Schmettow, M., Chamberlain, A., & Federici, S. (2023). Ciao AI: the Italian adaptation and validation of the Chatbot Usability Scale. *Personal and Ubiquitous Computing*, *27*(6), 2161–2170. https://doi.org/10.1007/s00779-023-01731-2

Brooke, J. (1996). SUS - A quick and dirty usability scale. In *Usability Evaluation in Industry* (Number 194, pp. 189–194). Taylor & Francis.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Digital.gov. (2025). *Usability | Digital.gov*. https://digital.gov/topics/usability/

Dix, A., Finlay, J., Abowd, G., & Beale, R. (2003). *Human-Computer Interaction (3rd Edition)* (3rd ed.). Prentice-Hall.

George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 Update* (4th ed.). Allyn & Bacon.

Hadiyati, R., & Tovtora S, F. D. D. (2025). The Impact of AI Chatbots, Service Personalization, and Response Speed on Customer Satisfaction in E-Commerce in Indonesia. *West Science Interdisciplinary Studies*, *3*(12), 2265–2275. https://doi.org/10.58812/wsis.v3i12.2497

Hambleton, R. K. (2005). *Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures* (R. K. Hambleton, P. F. Merenda, & C. D. Spielberger, Eds.; pp. 3–38). Lawrence Erlbaum Associates PublishersPress. https://www.taylorfrancis.com/books/9781135676575

Hasibuan, D. P., Santoso, H. B., Yunita, A., & Rahmah, A. (2020). An Indonesian Adaptation of the E-Learning Usability Scale. *Journal of Physics: Conference Series*, *1566*(1), 012051. https://doi.org/10.1088/1742-6596/1566/1/012051

Heffner, C. C., Fuhrmeister, P., Luthra, S., Mechtenberg, H., Saltzman, D., & Myers, E. B. (2022). Reliability and validity for perceptual flexibility in speech. *Brain and Language*, *226*, 105070. https://doi.org/10.1016/j.bandl.2021.105070

Hidayat, A., Nugroho, A., & Nurfaizin, S. (2022). Usability Evaluation on Educational Chatbot Using the System Usability Scale (SUS). *2022 Seventh International Conference on Informatics and Computing (ICIC)*, 01–05. https://doi.org/10.1109/ICIC56845.2022.10006991

Huuck, R. (2013). Formal Verification, Engineering and Business Value. *Electronic Proceedings in Theoretical Computer Science, EPTCS*, *105*, 1–4. https://doi.org/10.4204/EPTCS.105.1

Kuric, E., Demcak, P., & Krajcovic, M. (2025). Unmoderated Usability Studies Evolved: Can GPT Ask Useful Follow-up Questions? *International Journal of Human–Computer Interaction*, *41*(15), 9752–9769. https://doi.org/10.1080/10447318.2024.2427978

Mochammad Aldi Kushendriawan, Harry Budi Santoso, Panca O. Hadi Putra, & Martin Schrepp. (2021). Evaluating User Experience of a Mobile Health Application 'Halodoc' using User Experience Questionnaire and Usability Testing. *Jurnal Sistem Informasi*, *17*(1), 58–71. https://doi.org/10.21609/jsi.v17i1.1063

Nielsen, J. (1994). *Usability Engineering | Enhanced Reader*.

Noviyanti, C. E., Santoso, H. B., & Hadi Putra, P. O. (2021). A Cross-Cultural Adaptation of Chatbot Usability Questionnaire (CUQ): Indonesian Version. *2021 4th International Conference on Information and Communications Technology (ICOIACT)*, 248–251. https://doi.org/10.1109/ICOIACT53268.2021.9563926

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). McGraw-Hill.

Poglitsch, C., Seiser, M., Buchsteiner, M., & Pirker, J. (2025). AI Agents: Design and Evaluation of Gamified Conversational Agents. *2025 IEEE Conference on Games (CoG)*, 1–8. https://doi.org/10.1109/CoG64752.2025.11114224

Ren, R., Zapata, M., Castro, J. W., Dieste, O., & Acuna, S. T. (2022). Experimentation for Chatbot Usability Evaluation: A Secondary Study. *IEEE Access*, *10*, 12430–12464. https://doi.org/10.1109/ACCESS.2022.3145323

Rubin, J., & Chisnell, D. (2008). Handbook Of Usability Testing 2nd Ed. *Handbook Of Usability Testing 2nd Ed*, 366. https://www.wiley.com/en-us/Handbook+of+Usability+Testing%3A+How+to+Plan%2C+Design%2C+and+Conduct+Effective+Tests%2C+2nd+Edition-p-9780470185483

Santoso, H. B., Schrepp, M., Hasani, L. M., Fitriansyah, R., & Setyanto, A. (2022). The use of User Experience Questionnaire Plus (UEQ+) for cross-cultural UX research: evaluating Zoom and

Learn Quran Tajwid as online learning tools. *Heliyon*, *8*(11), e11748. https://doi.org/10.1016/j.heliyon.2022.e11748

Schrepp, M. (2020). On the Usage of Cronbach's Alpha to Measure Reliability of UX Scales. *Journal of Usability Studies*, *15*, 247–258.

Sharfina, Z., & Santoso, H. B. (2016). An Indonesian adaptation of the System Usability Scale (SUS). *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 145–148. https://doi.org/10.1109/ICACSIS.2016.7872776

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3205040

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tsakiridis, N. L., Diamantopoulos, T., Symeonidis, A. L., Theocharis, J. B., Iossifides, A., Chatzimisios, P., Pratos, G., & Kouvas, D. (2020). *Versatile Internet of Things for Agriculture: An eXplainable AI Approach* (pp. 180–191). https://doi.org/10.1007/978-3-030-49186-4_16

Tullis, T., & Albert, B. (2013). Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics: Second Edition. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics: Second Edition*, 1–301. https://doi.org/10.1016/C2011-0-00016-9

How to cite:

Santoso, H. R., Budiarti, Schrepp, M., Suryani, M., Aji, R. F., Nadine, E. J., & Dony, A. C. (2026). A Indonesian Adaptation of the Bot Usability Scale Full Version (BUS-42). *Jurnal Sistem Informasi (Journal of Information System)*, *22*(1), 33–47.