SN Computer Science (2021) 2:28
https://doi.org/10.1007/s42979-020-00424-4

ORIGINAL RESEARCH

Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa

Dilawar Shah Zwakman[1] · Debajyoti Pal[1] · Chonlameth Arpnikanondt[1]

Received: 8 October 2020 / Accepted: 10 December 2020 / Published online: 11 January 2021

Abstract

Currently, the use of voice-assistants has been on the rise, but a user-centric usability evaluation of these devices is a must for ensuring their success. System Usability Scale (SUS) is one such popular usability instrument in a Graphical User Interface (GUI) scenario. However, there are certain fundamental differences between GUI and voice-based systems, which makes it uncertain regarding the suitability of SUS in a voice scenario. The present work has a twofold objective: to check the suitability of SUS for usability evaluation of voice-assistants and developing a subjective scale in line with SUS that considers the unique aspects of voice-based communication. We call this scale as the Voice Usability Scale (VUS). For fulfilling the objectives, a subjective test is conducted with 62 participants. An Exploratory Factor Analysis suggests that SUS has a number of drawbacks for measuring the voice usability. Moreover, in case of VUS, the most optimal factor structure identifies three main components: usability, affective, and recognizability and visibility. The current findings should provide an initial starting point to form a useful theoretical and practical basis for subjective usability assessment of voice-based systems.

Keywords Graphical user interface · Voice-assistants · Usability · System usability scale · Voice usability scale · Factor analysis

Introduction

The use of voice-assistants has been on the rise in recent years. The global voice-assistant market was valued at 11.9 billion US dollars as of 2019 and projected to increase to 35.5 billion US dollars by 2025 [1]. Amazon *Alexa*, Google *Assistant*, *Siri* from Apple, and Samsung *Bixby* are some of the most popular voice-assistants available worldwide. In this study, voice-assistants are defined as "hardware devices or software agents that are powered by artificial intelligence and assist people with information searches, decision making efforts or executing certain tasks using natural language in a spoken format" [2]. The popularity of the voice-assistants is because of their ability to facilitate human–computer interactions in a natural and intuitive way, similar to the conversations between human beings. A recent survey indicated that 52% of the people prefer using a voice-assistant over a website for information search, because they found the voice-assistants more convenient [3]. Another 48% reported that they use voice-assistants as it allows them to multitask and work hands-free. Current academic literatures indicate the utilitarian as well as the hedonic benefits of voice-assistants along with their usability aspect that are critical factors for users to adopt and utilize these devices [4, 5].

Usability is an important aspect in the human–computer interaction domain that has attracted a lot of research interest over time. This is primarily because a good usability is an important performance measure of a product that is directly linked with the end-user satisfaction [6]. Improving the end-user satisfaction leads to a greater adoption of the products, and has been verified by well-established theories like Expectation Confirmation Theory [7], Expectation Confirmation Model [8], SERVQUAL Model [9], and various others. However, research regarding the measurement of usability of the voice-assistants and speech-based interfaces in general is lagging. Currently, there are no standard or well-defined metrics for measuring the usability of speech-based systems [10, 11]. The lack of any standardized measure makes it difficult to meaningfully assess these systems. The current work tries to address this gap in existing literature by considering two metrics: System Usability

✉ Debajyoti Pal
debajyoti.pal@gmail.com; debajyoti.pal@mail.kmutt.ac.th

[1] School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

Scale (SUS) and a new Voice Usability Scale (VUS) that is proposed in this work for assessing the usability with Amazon's *Alexa*.

The contribution of this work is twofold. The first research question that is investigated is "Is SUS a reliable and valid measure of usability for voice-assistants?". SUS is one of the most popular questionnaires used for the purpose of measuring usability in a graphical user interface (GUI) environment. Extant research has shown that SUS has an excellent reliability (typically the alpha coefficient exceeds 0.90), validity, and sensitive to a wide variety of independent variables [12, 13]. Thus, the efficacy of SUS as a usability measuring tool is well established. However, the problem is that the design of SUS instrument is heavily GUI oriented. Extant research has shown that voice-interfaces are distinctly different from GUIs and they have certain unique issues [11, 14]. For example, the ability to understand non-conversational cues (i.e., pauses in the middle of a conversation) [15], difficulty with back and forth navigation [15], and absence of a visual feedback that increases the cognitive workload [16] are some of the issues unique to a voice-based system. Therefore, how well SUS adapts to such a scenario is still an unexplored research problem. This is the first objective of this work.

Considering the uniqueness of voice-based systems, as the second objective, a new scale is developed called the VUS targeted toward the commercially available voice-assistants. While developing the new usability measurement tool, we had the option of selecting between an objective and subjective approach. The objective approach uses metrics like the task completion time, number of errors in doing a task, or measuring the physiological changes in the user when using the target product/system (e.g., the variation in heart-rate). On the contrary, subjective approach uses techniques like open interviews, focus groups, or questionnaires for gathering the information from the users after they have used the target product/system. One of the main advantages of this technique is that it is possible to get a variety of rich information and insights in different aspects of system acceptance that cannot be predicted before collecting the data. Moreover, since SUS is a 10-item questionnaire, therefore, to keep consistency, we decided to follow the subjective approach. The VUS scale is tested thoroughly in terms of its psychometric properties by doing the necessary reliability and validity analysis. Analyzing both these questionnaires in parallel not only enables to get an idea as to the suitability of SUS in the unique voice-only context, but also help to compare between the two instruments and decide which one captures the usability scenario better.

In the section "Literature review", we review the previous work in usability evaluation with special reference to SUS along with the current state-of-art of the voice-assistants. The section "Research methodology" presents the methodology of the study detailing the participants, questionnaire used and the experiment design. The data analysis and results are presented in the section "Data analysis and results". The section "Discussion and implications" provides the discussion along with the implications of the results. Finally, the section "Conclusion" presents the conclusion and the direction of future work.

Literature Review

This section is centered around the following three themes. First, the current state-of-art is presented in relation to the usability and adoption of the voice-assistants. Second, the SUS instrument is discussed in detail and why it is considered in this work. Third, the approach that is undertaken in this work based on the research gaps is presented.

Current State-of-Art of Voice-Assistants

Current research on voice-assistants stress on three main aspects: (a) improving the technology empowering these devices with an aim to provide better voice recognition, ability to understand multiple languages, providing human-like speech output, adding emotions to these devices, and likewise, (b) improving the privacy and security of these devices, so that the users can trust them in their daily usage, and (c) theoretical research focusing on the factors along with research models that explain the usage of these devices [17]. Although the indirect objective of the first research direction is to improve the usability of the voice-assistants by making them more user-friendly, yet these works focus on improving the technology, and not on the usability aspect per se. For example, authors in [18] develop a new system for maximizing the accuracy of an automatic speech-recognition engine by adjusting the front-end speech enhancement. They use a genetic algorithm to generate parameter values depending upon particular environments for improving the speech recognition. Researchers in [19] show how a semantically rich knowledge graph can be used to solve automatic speech recognition and language processing-specific problems. Using knowledge graph-based structured data, they build a unified system combining speech recognition and language understanding. A dynamic fusion framework combining empirical features and spectrogram-based statistical features is proposed by authors in [20] together with using a kernel extreme learning machine classifier for distinguishing emotions on two public speech emotion databases. Authors in [21] have utilized an Amazon Echo voice-assistant together with an ultrasonic sensor for detecting the location of elderly people in a smart-home environment. The incorporation of the voice-based features helps in reducing the burden of the learning curve of new technologies on family and caregivers,

thereby improving the quality of life. A similar scheme is proposed by authors in [22], where they use a voice-assistant along with a camera system for fall detection in a smart-home environment. Similar such works have been done by other authors in [23–25] for improving the various technical aspects relevant in voice-based systems.

The second research direction is with respect to the privacy and security aspect of the voice-assistants. These devices are prone to a variety of attacks that might steal user information. The security threats to the voice-assistants are not only potential, but very real and more dangerous as it is augmented by the inherent mechanisms of the underlying operating systems [26]. The authors in [26] demonstrate how various attacks can be launched on voice-assistants, along with their impact in real-world scenarios. Similar work is done by authors in [27] where they propose a novel attack vector named "Vaspy", which crafts the users' activation voice by silently listening to the users' phone calls. The attack vector is implemented with a proof-of-concept spyware and tested on different voice-assistants. Authors in [28] do a vulnerability analysis on different voice-based products like Google *Home*, Amazon *Alexa*, etc., and show that these devices can be exploited in multi-hop scenarios to maliciously access other Internet-of-Things (IoT) devices to which they may be connected. The replay attacks are modeled in a non-linear fashion that introduces higher order harmonic distortions. Authors in [29] design a completely inaudible attack that modulates voice commands on ultrasonic carriers that cannot be perceived by human beings. They validate the attack on a number of commercially available voice-assistants like Google *Home*, Amazon *Alexa*, and Microsoft *Cortana* by injecting a sequence of inaudible voice commands that leads to certain undesired actions. Authors in [30] propose a framework for an IoT home-security system that is secure, expandable, and accessible by integrating the system with an Amazon Echo voice-assistant. Therefore, from the above discussion, it is evident that considerable research efforts are being given to expose the existing vulnerabilities of the voice-assistants and methods to mitigate them for improving the trust of the users.

The third research dimension focus on identifying factors and building theoretical frameworks for explaining the adoption and usage scenario of the voice-assistants. A Uses and Gratification-based approach is taken by authors in [4] for understanding the motivations for adopting and using in-home voice-assistants (Amazon *Alexa*). The findings illustrate that individuals are motivated by the utilitarian benefits and symbolic benefits. The effect of hedonic benefit is applicable for small households only. Authors in [5] do an in-depth comparison of three popular adoption models (Theory of Planned Behavior, Technology Adoption Model, and Value-based Adoption Model) with respect to their capabilities of predicting the voice-assistant usage, and find Value-based Adoption Model to have the greatest predictive power. Additionally, a multiple regression analysis shows that the hedonic benefits of using the voice-assistants outweigh the utilitarian benefits. Authors in [31] use an extension of the Wixom & Todd's Information System Success Model to investigate extrinsic motivational factors that explain the voice-assistant usage scenario. Their findings suggest that trust, perceived risks, perceived enjoyment, and mobile self-efficacy affect the continuance usage of the voice-assistants. A comprehensive research model is proposed by authors in [32] based on the Perceived Value Theory, and it is found that perceived usefulness and enjoyment have a significant impact on the usage intention. A simultaneous acceptance and diffusion analysis of voice-assistants is done by authors in [33] for exploring the usage scenario of these devices. While the acceptance is explained by Technology Acceptance Model, the diffusion scenario is portrayed by the Multivariate Probit Model. Results show that usefulness, ease of use, compatibility, and perceived complementarity have significant positive effects on the purchase intention. In terms of the diffusion analysis the senior customers are more likely to purchase the voice-assistants within a given time frame when compared to the young customers. Although the theoretical works presented above try explaining the factors affecting the usage of the voice-assistants, yet such perceptions do not reflect the usability of these devices. No doubt that a better usability will translate to a better user experience, however, how to evaluate usability in the context of voice-assistants is a question that is not answered by the current theoretical works.

Research evaluating the usability of the voice-assistants is far from few. Authors in [34] develop a set of heuristics for assessing speech interfaces. They developed the heuristics based on existing principles of Nielsen and Molich's general heuristics for user interface design. However, the problem with such heuristics is that from a design perspective, they are deeply rooted in a GUI based environment. Authors in [35] compare four different voice-assistants (*Alexa,* Google *Assistant, Siri*, and *Cortana*) across four categories (shopping, travel and entertainment, administrative assistant and miscellaneous). Based on the correctness of the responses, the users give a rating on a scale of 1–5. While this work provides an initial direction to the usability assessment, however no standard questionnaires are used, neither the reliability nor validity of the questions that users ask to the voice-assistants are reported. This limits the scientific rigor of the work. A user-based summative usability testing of *Siri* is done by authors in [36] where the participants performed seven different tasks. After finishing the tasks, the participants had to fill up a short port-test questionnaire based on SUS. Although they used SUS as a standardized questionnaire, however, no empirical findings are reported by the work. Therefore, how good or bad SUS may represent the

Table 1 Summary of usability related studies of voice-assistants

Study no.	Usability approach	Sample size	Drawbacks
[34]	Heuristics based on Nielsen and Molich	16 participants (2 groups of 8 each)	1. Heuristic evaluation is done by experts 2. Typically, this method is used during early product lifecycle 3. How the participants interact with the voice-assistants is not elaborated
[35]	Non-standardized tool	8 participants	1. Number of participants too low to come to valid statistical conclusions 2. No standardized questionnaire is used 3. Results only provide descriptive statistics
[36]	SUS	20 participants	1. Results only provide descriptive statistics 2. No factor loadings of different SUS items are reported, neither their correlation 3. Whether SUS could capture the usability issues of voice-assistants in not clear
[37]	SUS	52 participants	1. Results only provide descriptive statistics 2. No factor loadings of different SUS items are reported 3. The tasks that users had to do with the voice-assistants is not clearly mentioned 4. No empirical evidence and discussion toward the suitability of SUS in measuring usability of voice-assistants
[38]	SUS	12 participants	1. The factor structure of SUS is not mentioned 2. The learnability dimension of SUS is non-significant indicating potential loading problems as per the original version 3. Differential loadings of the SUS items indicate that the voice scenario is different from GUI's 4. High correlation of SUS score with some other scale does not imply that SUS is a good usability measure for voice-assistants
[39]	Non-standardized tool	1462 survey participants	1. The exact nature of voice-assistants evaluated is not mentioned 2. The data is collected from survey with no mentions of whether the participants actually use a voice-assistant or not 3. How the participants interact with the voice-assistants is not mentioned that can produce biased results

usability scenario of the voice-assistants cannot be answered. Authors in [37] test *Siri* and *Alexa* with respect to their ability to understand commands from native and non-native English language speakers for performing certain common tasks. Results suggest no significant differences in usability between the two user groups. In this study, SUS is used for assessing the usability; however, the focus is more on the ability of the voice-assistants to understand the instructions from native or non-native English language speakers. Usability of *Alexa* and *Siri* is investigated by authors in [38] using the SUS questionnaire. They suggest that SUS is a valid evaluation tool for voice-assistants based on its strong correlation with the Adjective Rating Scale ($r=0.94$). However, we feel that such a judgment is too strong and early to conclude due to the following reasons. First, the authors did not publish the factor loadings of the different SUS items making it unclear about the relative magnitude of impact of the different items on the usability aspect. Second, the authors report some problems with items 4 and 10 (the learnability dimension of SUS) as, for these two items, the results are not-significant. It indicates that either SUS is not capable of handling the voice-based scenario well, especially the learnability aspect or the methodology of the study has certain limitations that biased the results. The differential loading of the SUS items into different dimensions is an indication that the original SUS structure or results might need certain modifications in a voice-based environment. In Table 1, a brief comparison of the various usability studies focusing on the voice-assistants is presented highlighting the drawbacks of the current works.

The System Usability Scale (SUS)

SUS was developed by Brooke [40] as a quick way to measure usability by the practitioners of a variety of products or service. Since its inception SUS has been very popular in the HCI community for assessing the perceived usability. SUS contains 10 items of mixed tone. Half of the items have a positive tone (the odd numbers), while the other half a negative tone (the even numbers). All the responses are taken on a scale of 1 (strongly disagree) to 5 (strongly agree). The final SUS scores range between 0 and 100, where a higher score means a better usability. Apart from SUS, a number of other usability evaluation tools are also available, for example

The Usability Metric for User Experience (UMUX), The Computer System Usability Questionnaire (CSUQ), Post-Study System Usability Questionnaire (PSSUQ), Software Usability Measurement Inventory (SUMI), and several others. In this work, we choose SUS due to the following main reasons:

(i) SUS is a widely used standardized questionnaire that is available free of cost. Authors in [13] reported that 43% of industrial usability studies use SUS. In fact, as of the current writing the original SUS work in [40] has been cited 9516 times that indicates its huge popularity.
(ii) Research has shown that SUS has excellent reliability, validity, and sensitivity to a wide variety of independent variables [41].
(iii) SUS questionnaire is simple and it provides a single score on a scale that is easily understood by a wide range of people (from project managers, computer programmers to normal end-users) who may have little or no experience in human factors engineering and usability.
(iv) SUS is technology agnostic, therefore making it a flexible tool for assessing usability of a variety of technologies, from traditional computer interfaces to websites and even software products [13, 41].

The excellent psychometric properties of SUS together with the reasons mentioned above make it a good candidate for usability evaluation. However, there are a few areas of concern. First, although SUS is technology agnostic, still, its use has been primarily to evaluate the usability of GUI-based systems. Voice-assistants have come into prominence very recently and they have a number of unique features that are not addressed by usability evaluation scales targeted toward GUI-based systems. For example, speech recognition is an important aspect of voice-assistants, and despite improvements in speech recognition technology, they make errors [36, 37]. Consequently, they must give feedback to the users as to what has been recognized. The question of how accurate the speech recognition of the voice-assistants must be, while still being useful and acceptable to the users is a crucial factor for their success. Moreover, based on the experiences of human-to-human conversations, the users have some strong pre-conceived and obvious expectations from the voice-assistants as to how a conversation should proceed. Often times during a natural human conversation, there is a pause between different words for the sake of clear understanding. However, speech-based systems have a difficulty in understanding such non-conversational cues (pauses in the middle of a conversation) [15]. These aspects of naturalness or intuitiveness during conversations are important that makes these systems unique and different from GUIs. Another aspect of speech-based systems is their lack of visual feedback. Although some recent voice-assistants available in the market are now coming with screens, however, such an additional modality is an exception, and not a norm. Lack of any type of visual feedback has been seen to increase the cognitive load of the users [42], which might have an effect on their usability. In the absence of any form of visual menu or guides, it might be difficult for the users to learn using the system or apprehend what the system does. This is another unique feature of the voice-only scenario that is different from the GUIs. Extant research also shows that quality of the synthetic voice may affect the perception of the users. For example, authors in [43, 44] asked the participants to evaluate the naturalness and intelligibility of certain speech outputs from machines using various scales like the Absolute Category Rating (ACR) and Degradation Category Rating (DCR) scales, and found that the speech quality impacts the users satisfaction levels. The voice-assistants are anthropomorphic devices, and several

Table 2 Assessment scheme for the proposed VUS usability instrument

Usability dimension	Explanation
General/usability	There is no specific focus on any aspect. It captures the general impression or sentiments of the users after using the voice-assistants
Affective	The psychological state of the users (emotions/feelings/impressions) after using the voice-assistants. The users' feelings (happy/pleased/satisfied) are reflected by this dimension
Recognition & visibility	Users must recognize the various functions and options just through interaction and affordance with the voice-assistants. The voice-assistants must provide interaction in a natural and intuitive manner rather than stating what kinds of commands someone can give
Pragmatic	Ability of the voice-assistants to support 'goal-oriented' tasks (e.g., making a call, searching for information, etc.). The users will mainly judge the efficiency/usefulness of the voice-assistants
Errors & frustration	The voice-assistants should have constraints built in place to help users not to come across errors. Cascading errors should be avoided. In the event of some error, the voice-assistants must allow the users to exit from errors or a mistaken conversation
Guidance & help	The voice-assistants must provide guidance to the users through their interactions, so that they are not easily lost. Interaction must be short for minimizing the acoustic confusability of vocabulary (i.e., short yes/no type)

studies have revealed that users consider them to be their assistive companions [2, 4, 17, 45]. Therefore, quality of the auditory feedback is an important issue that is another uniqueness of the voice-based systems.

From the above discussion, it is imperative that voice-assistants have many features that are unique and not found in GUI-based environments. Moreover, although SUS is robust, yet it is primarily meant for testing GUI-based systems. The voice-assistants have come into market recently; therefore, whether standardized instruments like SUS can be used for measuring their usability is unknown. However, considering the radical differences between voice-interfaces and GUIs in this work, we propose the development of a new scale targeted toward the voice-assistants called the Voice Usability Scale (VUS). The next section introduces this new scale.

The Development of Voice Usability Scale (VUS)

Given the limitations of SUS in a voice-based scenario, there is a need for the development of a more valid and reliable approach for the present context. Drawing parallels to SUS, the proposed VUS scale is also a 10-item one allowing meaningful comparisons to be made between the two. The problem of developing a usability measure for voice-assistants is that there are no commonly accepted usability dimensions. Recently, Murad et al. [15] and Wei, Landay [46] proposed some HCI design guidelines for voice-based smart devices based on a heuristic approach. Although heuristic evaluation and usability testing are two totally different approaches of usability evaluation, yet they share one common goal, i.e., evaluating the usability. Table 2 illustrates the assessment scheme for the VUS measurement instrument. As evident from the table, we try to capture different aspects of usability as applicable to the current context of voice-assistants. In this aspect, it is important to note that the usability

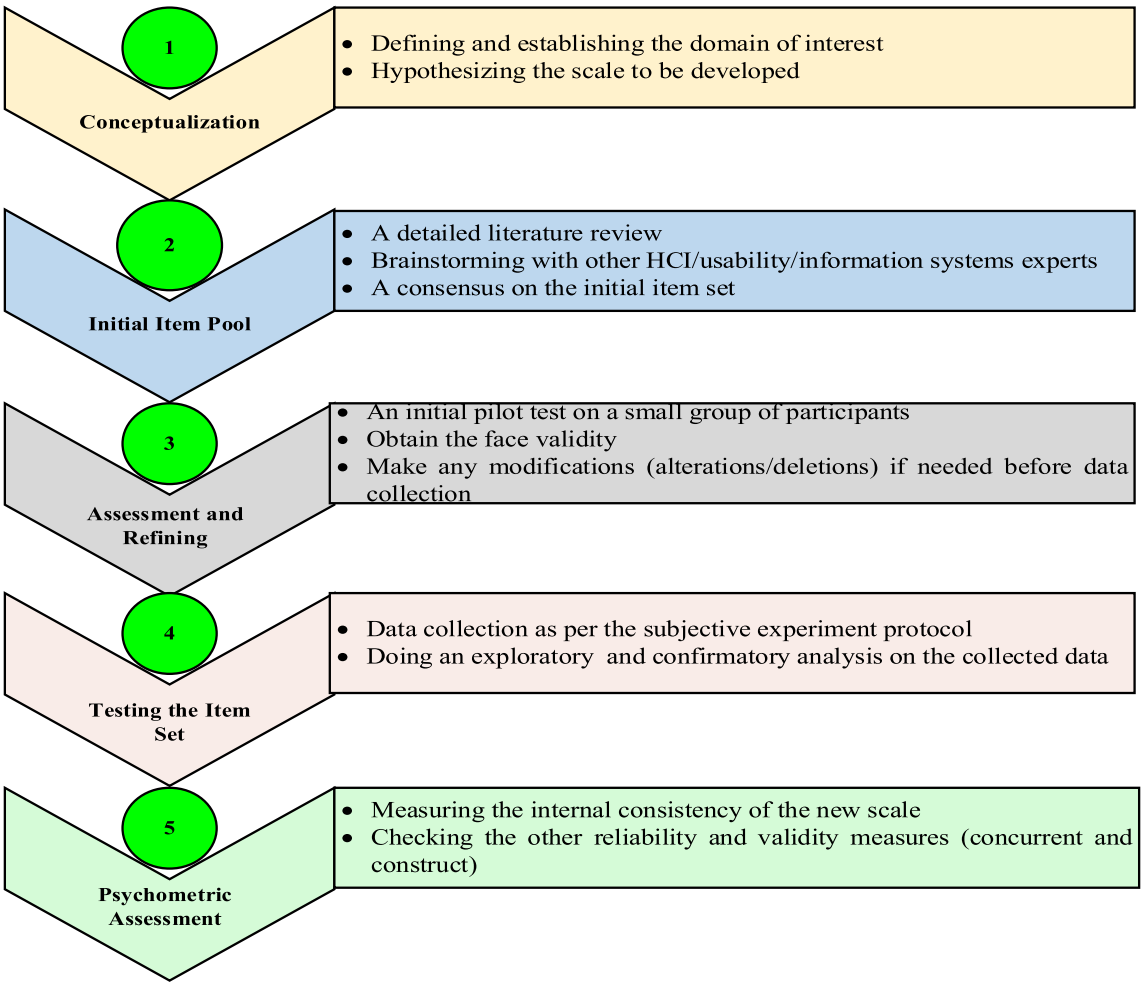

Fig. 1 Flowchart of the overall methodology

dimensions that are identified in Table 2 are not complete; rather, they are representative and indicative. Usability evaluation of voice-assistants is an emerging area, and our objective is not to exhaustively evaluate the usability dimensions, rather identify the common usability aspects, and use them as a reference to understand the overall scope and characteristics of usability assessment for this novel scenario.

Due to the absence of firm theoretical guidance related to the features of the voice-assistants that effect the usability, we decided to use an empirical approach for developing the VUS instrument. Toward this goal, initially based on the literatures just discussed (especially [15] and [46]) along with discussing with other HCI experts, a pool of items is first generated. A pilot study is conducted on this initial item-set. Based on the feedbacks received, some items are either modified or deleted. This is followed by a comprehensive usability testing by participants on the revised instrument. The current work describes the first in a series of planned iterations for designing the VUS. The detailed research methodology is outlined in the next section.

Research Methodology

To fulfill the dual objectives of this work in the first step, the items of the VUS questionnaire are generated. In the second step, the pilot-tested questionnaire is administered among the sample participants after using an Amazon *Alexa* smart-speaker. The overall flow of the experiment is depicted in Fig. 1.

VUS Item Generation

Right at the onset, it was decided to use a 7-point Likert scale having declarative statements of opinion to which the participants will respond with their rate of agreement. This method is chosen over other alternatives (for, e.g, a bipolar scale) due to the following reasons. First, we wanted the VUS scale to be similar in structure to SUS for the ease of comparison. Second, in case of a bipolar scale sometimes, it becomes very difficult to determine the appropriate opposites. Third, with declarative statements, it is possible to capture a finer grain of meaning from the different items. An initial set of items is generated based on the literature review of existing usability instruments that are discussed previously in the literature review section. While including these initial items, the authors' subjective opinions and practical experiences

Table 3 SUS and VUS questionnaires

Item	SUS	VUS
1	I think I would like to use the voice-assistants frequently	I thought the response from the voice-assistant was easy to understand
2	I found the voice-assistant unnecessarily complex	I thought the information provided by the voice-assistant was not relevant to what I asked
3	I though the voice-assistant was easy to use	My interaction with the voice-assistants was fast
4	I think that I would need the support of a technical person to use this voice-assistant	I thought the voice-assistant had difficulty in understanding what I asked it to do
5	I found the various functions in this voice-assistant were well integrated	I felt the voice-assistant enabled me to successfully complete my tasks when I required help
6	I thought that there was too much inconsistency in this voice-assistant	It was easy to lose track of where you were in an interaction with the voice-assistants
7	I imagine that most people would learn to use this voice-assistant very quickly	The voice-assistant had all the functions and capabilities that I expected it to have
8	I found the voice-assistant very awkward to use	I found it difficult to customize the voice-assistant according to my needs and preferences
9	I felt very confident using the voice-assistant	Overall, I am satisfied with using the voice-assistant
10	I needed to learn a lot of things before I could get going with this voice-assistant	I found the voice-assistant difficult to use
11	×	I felt the response from the voice-assistant was sufficient
12	×	I found it frustrating to use the voice-assistant in a noisy and loud environment
13	×	I was able to recover easily from errors
14	×	The voice-assistant was unreliable

are taken into consideration. Care is taken to balance the positive and negative statements.

After generating the initial items, they are shown to 6 experts (2 each from HCI, usability, and information systems domain). The experts mainly focus on checking the clarity of meaning of the included items, and all the confusing items are removed. If there is any disagreement regarding the inclusion/exclusion of any item, then a majority vote is taken. In the event of a tie the item is included. By following this procedure of item generation, an initial pool of 15 items are generated. The pilot testing is done on a small sample of 14 participants. All the participants are recruited from the authors' university and they had experience in using voice-assistants before. One item is removed after the pilot test, because it could not be generalized to all types of voice-based systems and had a low face validity. The 10-item SUS scale and the proposed 14 item VUS scale are elaborated in Table 3.

Sample and Experiment Protocol

The participants are recruited both from the authors' university and outside. The internal participants are contacted through mailing-lists and personal contacts, whereas those from outside are contacted through various social-media channels (Facebook, Line, and WhatsApp). The choice of selecting Amazon Echo as the smart-speaker is made due to the following reasons. First, Amazon is the current market leader in voice-assistants. As of 2020 the global market share of Alexa stands at 31.7%. Second, the voice-assistant platform provided by Amazon is matured and advanced, supporting a variety of skills and routines enabling the users to perform a multitude of tasks. Moreover, majority of the smart-home products and accessories are compatible with the devices available from Amazon. In the experiment, the 3rd generation of Echo Dot speaker is used. The experiment has passed the Institutional Review Board (IRB) committee of the authors' university prior to its administration on the participants.

Initially 88 participants were selected for the experiment, although 62 take part finally due to an exclusion criterion used. The requirements for participating in the experiment was having a good command over English and being at least 18 years of age. All the participants had good English proficiency skills in terms of their IELTS, TOEIC, TEOFL iBT (Internet-based), or TETET scores. From the final analysis, one participant is excluded (final sample size = 61) as it is an extreme outlier. The participant selected the same value for all the items in both the questionnaires, which is either the maximum value or the minimum value. The percentage of male and female participants is roughly equal with a mean age of 24.5 years (SD = 5.49). All the participants have previous experience in using some form of voice-assistant that makes them well-suited to understand the purpose behind the experiment and reduce the response bias as such. Before the start of the experiment as per the guidelines of any standardized evaluation procedure, clear instructions are provided to the participants: "Thank you for agreeing to take part in this evaluation. It will take no longer than 15 to 20 minutes to complete the whole process. First, you will have to ask some questions to the smart speaker that is kept in front of you (check the supplied printout for the question details). Based on your experience with the smart-speaker please rate the items that follow the questions in the printout. Please note that this is not a test of you—you are simply using the questionnaire and your experience with this smart-speaker to obtain a general perception of its usability. Please read and mark each item carefully. Your first impression is just fine."

To begin with the overall objective and the general procedure of the experiment is explained to the participants.

Table 4 Overview for some of the questions

Question no.	Questions
1	What is the weather like in Bangkok? How about tomorrow? Will it rain on Friday?
2	What is todays date? What time is it now? Why is the sky blue?
3	What is COVID-19?
4	What do you know about Asia? What about Nobel Prize?
5	Add milk to my shopping list. Add eggs and jam to my shopping list
6	How many eggs did you add on my shopping list?
7	What is on my shopping list?
8	Set an alarm for 08:30
9	I need to make an appointment with doctor
10	Set a schedule for 12:30
11	How can I protect myself from corona virus?
12	Give me some words of wisdom
13	What is 10 plus 5? Add 20 to the result. What is the final result? Divide 20 by 0
14	How to go to Siam BTS? How much time will it take?

Each of the participants are provided with a script that contains some questions which they need to ask to the voice-assistants. Before starting the experiment, the Amazon Echo Dot smart-speaker is set-up and configured properly with all the services that are needed for the purpose of evaluation. For example, the speaker is linked to a paid Amazon Prime account of one of the authors. A variety of questions are asked, so that most of the usability dimensions that are presented in Table 2 are covered. The questions ranged from simple ones like asking the current weather, date, and time, to more complicated scenarios like adding items to shopping list, making an appointment with the doctor and many more. The questions are framed in a manner, such that the high-level features and functionalities provided by the voice-assistants are revealed. The participants were informed that they are free to retry completing any tasks any number of times they want. Moreover, it was also informed to them that if they felt uncomfortable, they could quit from the experiment at any time, without fearing any negative consequences. An overview of some of the questions that the participants asked is presented in Table 4. The excerpt below related to one task (playing music) is shown in detail for a random participant.

PARTICIPANT 42: Hey Alexa, play Celine Dion.
ALEXA: Playing songs by Celine Dion from Amazon Music [*Starts playing a random song*].
PARTICIPANT 42: Alexa, Stop.
ALEXA: [Stops playing the song].
PARTICIPANT 42: Alexa, play a playlist Romantic Sundays.
ALEXA: What do you want to hear?
PARTICIPANT 42: Romantic Sundays.
ALEXA: I can't find a song Romantic Sundays.
PARTICIPANT 42: Alexa, play a playlist.
ALEXA: Ok, what do you want to hear?
PARTICIPANT 42: Nineties.
ALEXA: Sorry, I couldn't find any Nineties playlist.

After finishing the entire script, the participants had to fill up the SUS and VUS questionnaires on a computer. The order in which the questionnaires are presented is randomized to minimize bias. At the end, the participants had to answer an additional question based on their overall experience with the voice-assistant: "Overall, I would rate the user-friendliness of this voice-assistant as worst-imaginable/awful/poor/ok/good/excellent/best-imaginable". This question represents the Adjective Rating Scale [47]. The experiment lasted for about 35 min on an average for each participant. As a token of appreciation, small gifts are given to the participants who took part in the experiment.

Table 5 Distribution of the SUS and VUS scores

Parameters	Previous study [41]	SUS (present study)	VUS (present study)
N	2324	61	61
Minimum	0.00	38.32	45.00
Maximum	100.00	89.96	98.33
Mean	70.14	63.69	70.19
Median	75.00	63.31	69.99
Standard deviation	21.71	11.44	15.25
Standard error	0.45	1.46	1.95
First quartile	55.00	61.64	56.66
Third quartile	87.50	77.47	83.34
Inter-quartile range	32.50	15.83	26.68
99.9% confidence interval (upper)	71.50	65.76	75.38
99.9% confidence interval (lower)	68.70	58.62	64.99

Data Analysis and Results

Data Preprocessing

The data analysis is done in SPSS version 13.0. Before analyzing the data, it is checked for any missing values, the data distribution, and the assumptions of multivariate analysis. One item belonging to the VUS questionnaire (item number 3) had a lot of missing data (more than 20% of the respondents). Consequently, item 3 is removed from any further analysis. Apart from this specific case, another 42 missing data points were detected distributed randomly throughout the dataset. Since these points are randomly distributed with no pattern apparently, they are replaced with the mean values (calculated from the remaining cases) for each of the points. Next, the distribution of SUS and VUS scores is checked. For every participant, all the individual item rating (for both SUS and VUS) are first converted to a total score (between 0 and 100). For the odd numbered items (positively worded), the score contribution is 7 minus the item rating, whereas; for the even numbered items (negatively worded), the score contribution is the item rating minus 7. To get the overall score, the summation of the normalized item scores is multiplied by a constant factor of 1.667. Equation (1) shows the score calculation procedure:

$$\text{Score}_{\text{TOTAL}} = 1.667 \times \left(\sum \text{Score}_{\text{ODD}} + \sum \text{Score}_{\text{EVEN}} \right), \tag{1}$$

where; $\sum \text{Score}_{\text{ODD}} = \sum_{i=1}^{n-1} \left(7 - x_i\right)$, and $\sum \text{Score}_{\text{EVEN}} = \sum_{j=2}^{n} \left(y_j - 7\right)$, x_i denotes score of the odd items, y_j denotes the score of the even items, and n denotes the total number of items.

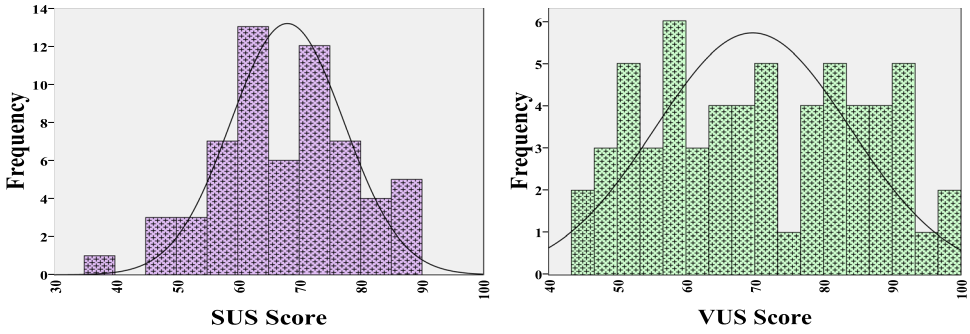

Fig. 2 SUS and VUS score distribution

Table 6 Relevant statistical measures for KMO and Bartlett's test

Statistic	SUS	VUS
KMO (sampling adequacy)	0.733	0.790
Bartlett's test Chi-square	185.01	252.738
Degree of freedom (*df*)	45	45
Significance	<0.001	<0.001

Authors in [41] provided a distribution of SUS scores collected from 2324 surveys over a course of 206 studies. The relevant descriptive statistics for the current context are presented in Table 5. A graphical distribution of the SUS and VUS scores are shown in Fig. 2. The skewness, kurtosis, and linearity of the data are also found to be satisfactory.

Next, the correlation matrix is examined to check whether the requirements of factor analysis are met. There are several correlations of 0.30 or higher, suggesting that the data are suitable for factor analysis. Furthermore, a Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy is done that indicates the proportion of variance in the different items that might be caused by underlying factors. The Bartlett's test of sphericity tests the hypothesis that the correlation matrix is an identity matrix, which indicates that the items are unrelated and hence suitable for structure determination. The relevant statistical results are reported in Table 6. For both cases, the KMO value is greater than the threshold of 0.5 and the Bartlett's test is significant indicating the suitability of doing a factor analysis [48].

Psychometric Properties and Reliability Analysis

Extant research has shown that SUS has excellent psychometric properties [13, 40, 41]. However, since the current context is regarding the usability of voice-interfaces, therefore it needs to be seen how SUS behaves presently. Typically, the reliability of SUS exceeds 0.90 (greater than the threshold limit of 0.70) as per the existing studies [40, 41, 47]. Before carrying out the reliability analysis, all the negatively worded items (the even numbered ones) are transformed to their absolute values, so that the entire scale is uniform (1 means the worst and 7 means the best). For the SUS and VUS scales, the values of coefficient alpha (Cronbach's α) are 0.773 and 0.807, respectively. In case of SUS, although the value is greater than the threshold, yet it is far below the previously reported results. Nevertheless, both the scales are reliable.

Exploratory Factor Analysis

Since both the SUS and VUS scales are designed to give a single score, an exploratory factor analysis (EFA) is carried out to determine whether the different items of the questionnaires measure the usability aspect as intended or something else. EFA is generally used in cases where there is a need to discover the structure of the collected data and to examine its internal reliability. This multivariate statistical technique is commonly used by researchers when developing a new scale, as is done in this work. In case of SUS, previous studies

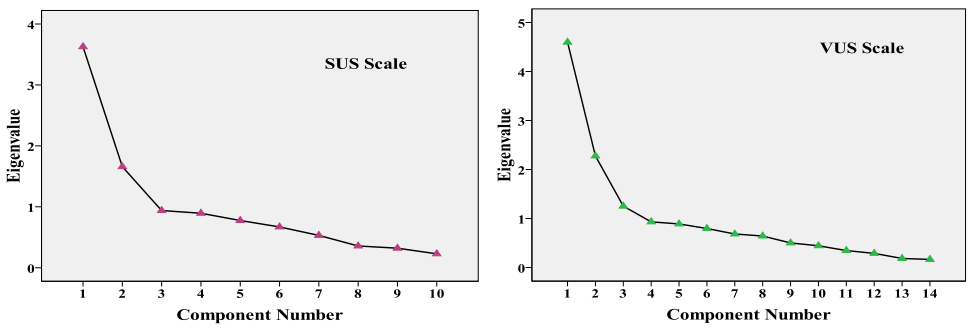

Fig. 3 Scree plot for SUS and VUS scales

Table 7 Factor analysis for SUS dataset (four- vs. three- vs. two-factor solutions)

Methodology	Items	Four Factor Solution				Three Factor Solution			Two Factor Solution	
		1	2	3	4	1	2	3	1	2
Principal Component Analysis	SUS_1	0.068	**0.855**	0.113	0.036	0.077	**0.830**	0.102	0.096	**0.757**
	SUS_2	**0.721**	-0.109	-0.163	-0.184	**0.719**	-0.133	-0.233	**0.723**	-0.240
	SUS_3	**0.848**	0.000	0.147	0.120	**0.847**	0.016	0.193	**0.853**	0.120
	SUS_4	**0.734**	-0.021	0.079	0.149	**0.733**	0.000	0.132	**0.728**	0.067
	SUS_5	-0.297	**0.723**	0.116	0.209	-0.292	**0.739**	0.189	-0.278	**0.727**
	SUS_6	**0.787**	-0.077	0.048	0.026	**0.787**	-0.078	0.049	**0.782**	-0.044
	SUS_7	-0.032	0.203	**0.955**	0.020	-0.015	0.118	**0.855**	-0.041	**0.454**
	SUS_8	**0.783**	-0.276	0.122	0.013	**0.782**	-0.281	0.113	**0.770**	-0.182
	SUS_9	-0.032	0.168	0.022	**0.963**	-0.040	0.361	**0.458**	-0.046	**0.449**
	SUS_{10}	**0.603**	0.285	-0.261	-0.169	**0.604**	0.262	-0.320	**0.621**	0.050
Unweighted Least Squares	SUS_1	No Minimum Solution is Found (25 Iterations)				-0.084	**0.624**	0.074	-0.086	**0.627**
	SUS_2					**0.667**	0.182	0.127	**0.664**	0.222
	SUS_3					**0.852**	0.092	0.112	**0.848**	0.130
	SUS_4					**0.643**	-0.001	-0.042	**0.646**	-0.014
	SUS_5					0.248	**0.699**	0.056	0.247	**0.687**
	SUS_6					**0.717**	0.104	-0.044	**0.720**	0.084
	SUS_7					0.049	0.254	**0.882**	0.058	**0.347**
	SUS_8					**0.720**	0.273	-0.118	**0.712**	0.220
	SUS_9					0.066	**0.317**	0.075	0.061	**0.348**
	SUS_{10}					**0.531**	-0.048	0.096	**0.527**	-0.011
Maximum Likelihood Analysis	SUS_1	-0.204	**0.883**	0.079	0.262	-0.248	**0.836**	0.405	-0.194	**0.966**
	SUS_2	**0.628**	0.155	0.134	0.239	**0.634**	0.183	0.213	**0.657**	0.214
	SUS_3	**0.879**	0.081	0.094	0.035	**0.882**	0.163	-0.019	**0.874**	0.083
	SUS_4	0.545	-0.056	-0.034	**0.835**	0.609	-0.131	**0.782**	**0.582**	0.170
	SUS_5	0.320	**0.549**	0.110	-0.078	0.278	**0.582**	-0.023	0.296	**0.481**
	SUS_6	**0.695**	0.088	-0.037	0.145	**0.691**	0.113	0.116	**0.706**	0.110
	SUS_7	0.069	0.203	**0.976**	-0.015	0.084	**0.325**	-0.036	0.090	**0.265**
	SUS_8	**0.716**	0.305	-0.121	0.063	**0.672**	0.314	0.080	**0.705**	0.279
	SUS_9	0.099	**0.273**	0.045	-0.071	0.083	**0.295**	-0.043	0.091	**0.233**
	SUS_{10}	**0.507**	-0.012	0.078	0.082	**0.513**	0.029	0.045	**0.521**	0.019

***Note:** The numbers marked in bold represent the highest loading on the factor

The numbers marked in bold represent the highest loading on the factor

report the presence of both one-factor as well as two-factor solutions [13, 41]. With regards to the EFA methodology, popularly, three techniques are used: Principal Component Analysis (PCA), Unweighted Least-squares Factor Analysis (ULS), and Maximum-Likelihood Factor Analysis (ML). For all the three techniques, an orthogonal rotation is used (varimax rotation algorithm). PCA is an unsupervised technique in which the interrelated items are transformed to a new set of items (called principal components) in such a way that they are uncorrelated and the first few of these components explain most of the variance of the entire dataset. The ULS technique aims at minimizing the residuals between the input correlation matrix and the reproduced correlation matrix. Finally, the ML technique is often used in cases where the input data are normally distributed, as it is in the present scenario (Fig. 2 gives an idea of the normal score distribution). Our strategy was to start with a four-factor solution and then work our way downwards to three- and two-factor solutions. The scree plot for the dataset (both SUS and VUS) is presented in Fig. 3. In case of SUS, there are 2 factors for which the eigenvalue is greater than 1, and in case of VUS, there are 3 factors for which the eigenvalue is greater than 1. However, the third and fourth factors (for SUS) and the fourth factor (for VUS) are marginally lower than 1. Therefore, comparing the two-, three-, and four-factor solutions is justified.

In case of the new VUS scale, an initial visual walkthrough of the factor analysis revealed several anomalies. First, some of the items did not load onto any factor, i.e., they had a very low loading. Although, extant literatures point toward an optimal loading of 0.4 below which the item should be deleted, we followed a stricter rule of 0.5. Similarly, for some of the items, there is a cross loading problem (either loading with values greater than 0.4 or the absolute difference in the magnitude of the loadings in less than 0.2) on multiple factors. Following this procedure, three items are eliminated (item numbers 6, 13, and 14). An inspection of these removed items shows some of them to be potentially ambiguous, therefore justifying their removal. The factor matrix for both SUS and VUS is presented in Tables 7 and 8, respectively. In case of SUS, the four-, three-, and two-factor solutions account for 71.18%, 62.24%, and 52.85% of the total variance, respectively.

Table 8 Factor analysis for VUS dataset (four- vs. three- vs. two-factor solutions)

Methodology	Items	Four-factor solution				Three-factor solution			Two-factor solution	
		1	2	3	4	1	2	3	1	2
Principal component analysis	VUS_1	0.106	0.115	**0.881**	0.115	− 0.007	0.303	**0.775**	0.108	**0.622**
	VUS_2	**0.856**	0.177	0.247	0.126	**0.829**	0.258	0.222	**0.852**	0.276
	VUS_4	**0.835**	0.263	0.132	0.209	**0.844**	0.280	0.192	**0.859**	0.290
	VUS_5	− 0.023	0.479	**0.633**	0.026	− 0.071	**0.715**	0.255	− 0.028	**0.757**
	VUS_7	0.186	**0.817**	0.181	− 0.262	0.203	**0.813**	− 0.317	0.113	**0.710**
	VUS_8	0.366	0.002	0.134	**0.788**	0.321	0.041	**0.688**	**0.560**	0.186
	VUS_9	0.323	**0.719**	0.226	0.162	0.366	**0.735**	0.064	0.356	**0.717**
	VUS_{10}	**0.827**	− 0.023	0.094	0.301	**0.824**	0.007	0.320	**0.874**	0.054
	VUS_{11}	− 0.080	**0.708**	0.049	0.532	0.037	**0.640**	0.228	0.073	**0.674**
	VUS_{12}	**0.730**	0.071	− 0.281	− 0.113	**0.752**	− 0.071	− 0.244	**0.682**	− 0.150
Unweighted least squares	VUS_1	0.101	0.126	**0.856**	0.063	0.115	0.038	**0.578**	0.117	0.502
	VUS_2	**0.849**	0.220	0.196	0.029	**0.839**	0.241	0.160	**0.843**	0.290
	VUS_4	**0.858**	0.237	0.125	0.121	**0.855**	0.234	0.197	**0.855**	0.309
	VUS_5	0.033	0.290	**0.425**	0.300	0.003	**0.655**	0.155	0.003	**0.643**
	VUS_7	0.079	**0.979**	0.136	0.130	0.079	0.361	**0.925**	0.109	**0.633**
	VUS_8	**0.465**	− 0.051	0.224	0.164	**0.558**	0.291	− 0.110	**0.549**	0.193
	VUS_9	0.333	**0.492**	0.318	0.273	0.221	**0.541**	0.301	0.310	**0.711**
	VUS_{10}	**0.858**	− 0.043	0.062	0.088	**0.852**	0.118	− 0.067	**0.854**	0.059
	VUS_{11}	0.083	0.196	0.148	**0.966**	0.091	**0.533**	0.365	0.085	0.560
	VUS_{12}	**0.537**	0.086	− 0.135	− 0.078	**0.550**	− 0.156	0.128	0.536	− 0.057
Maximum-likelihood analysis	VUS_1	0.092	**0.457**	0.204	0.217	0.106	0.039	**0.587**	0.134	**0.558**
	VUS_2	**0.854**	0.208	0.200	0.121	**0.853**	0.248	0.166	**0.854**	0.292
	VUS_4	**0.824**	0.151	0.265	0.196	**0.844**	0.253	0.186	**0.837**	0.326
	VUS_5	0.021	**0.882**	0.170	− 0.074	0.019	**0.624**	0.157	0.050	**0.546**
	VUS_7	0.133	0.264	**0.782**	− 0.118	0.084	0.367	**0.921**	0.091	**0.721**
	VUS_8	0.334	0.120	0.028	**0.689**	**0.539**	0.324	− 0.151	**0.648**	0.181
	VUS_9	0.233	0.311	**0.668**	0.318	0.285	**0.562**	0.495	0.267	**0.757**
	VUS_{10}	**0.857**	0.130	− 0.079	0.187	**0.861**	0.120	− 0.060	**0.880**	0.030
	VUS_{11}	0.053	**0.425**	0.305	0.211	0.076	**0.547**	0.152	0.093	**0.518**
	VUS_{12}	**0.517**	− 0.134	0.104	0.073	**0.541**	− 0.118	0.088	**0.513**	0.008

The numbers marked in bold represent the highest loading on the factor

Similarly, for VUS, the variance explained is 75.94%, 68.03%, and 58.18% for the four-, three-, and two-factor solutions, respectively. For both the cases, the communality values are acceptable (greater than or equal to 0.4). The results indicate that, for SUS, the optimal solution is obtained using the PCA algorithm for the two-component version (Table 7). For the other two algorithms, low loadings are obtained for some of the items. Only PCA gives a clean solution; however, the loadings of items 7 and 9 are 0.454 and 0.449, respectively, indicating that they are just above the acceptable limit of 0.4. For VUS, the obtained results are different. In this case, both the three-factor and two-factor solutions are valid (Table 8). Figure 3 also indicates the existence of 3 factors for the VUS scale. Since there is a tie between the three- and two-factor solutions, therefore, we decided to run a parallel analysis in addition to the eigenvector technique that is used. Parallel analysis is also a powerful technique based on Monte Carlo simulation that works by creating a random dataset with the same number of observations and variables as the original data. Parallel analysis also indicated the presence of three factors. Although, these statistical properties must be considered while doing a factor analysis, yet the interpretability of the results depends on the judgment of the analyst. In case of VUS items (2, 4, 10, 12, Factor 1), (5, 7, 9, 11, Factor 2), and (1, 8, Factor 3) loaded together with PCA giving the most optimal solution. Factor 1 contains items that bring out the errors and difficulties that users encounter while using the voice-assistants and consequently the frustrations, for example, "I thought the voice-assistant had difficulty in understanding what I asked it to do", "I

thought the information provided by the voice-assistant was not relevant to what I asked" and more. We named this factor as Usability. Factor 2 contain items like "Overall, I am satisfied with using the voice-assistant" and "I felt the voice-assistant enabled me to successfully complete my tasks when I required help" that are similar to how a user feels after using the voice-assistants. Thus, this factor is named as *Affective*. Finally, factor 3 contain items like "I thought the response from the voice-assistant was easy to understand" and "I found it difficult to customize the voice-assistant according to my needs and preferences". These items are clearly related to the fact that whether the voice-assistants recognize the users properly and give back a satisfactory response as expected, and whether they are easily customizable. Hence, we named this factor as *Recognizability & Visibility*. For all the three factors (sub-scales), the reliability measures are re-calculated and found to be acceptable: (1) *recognizability & visibility,* $\alpha = 0.85$, (2) *usability*, $\alpha = 0.82$, and (3) *affective*, $\alpha = 0.87$.

SUS, VUS, and Usability of Voice-Assistants

Over the years, since the Adjective Rating Scale (ARS) has been found to correlate well with SUS [41, 47], we decided to find the relationship between (ARS, SUS) and (ARS, VUS) to check how well they are related for the present scenario. For this, a correlational analysis is carried out (using the ARS scores matched with the corresponding SUS/VUS scores of the participants). In case of SUS, the results are significant ($p < 0.01$) with a low Pearson correlation value of 0.203. However, the same in case of VUS is high (0.765), the results still being significant ($p < 0.01$). For SUS, the obtained result is significantly different from previous studies [41, 47], indicating that it might not be a good measure of usability for the voice-only context. On the contrary VUS seems to be a good measure.

Discussion and Implications

Given the uniqueness of voice-assistants and their difference from GUI-based systems, this work investigates two questions. First, the suitability of SUS which is one of the most popular usability evaluation scales for GUI-based systems is checked for the voice-only context. Second, a new standardized scale is proposed keeping in mind the specific requirements of the voice-assistants.

SUS and the Usability of Voice-Assistants

SUS is a standard scale and an extremely popular tool for measuring the usability of GUI-based products. However, several new and different observations are obtained when using SUS for evaluating the usability of the voice-assistants. First, when comparing SUS scores from previous studies [41] with those currently obtained (Table 5), it is seen that the central tendencies of the distributions are not identical. The mean SUS score reported in [41] is 70.1 with a 99% confidence interval ranging from 68.7 to 71.5. However, presently, we obtain a mean SUS score of 63.69 with a 99% confidence interval ranging from 58.62 to 65.76. Therefore, the confidence intervals for the two cases are non-overlapping, and the difference in mean is statistically significant ($p < 0.01$).

The second major difference is with respect to the factor structure of SUS. Extant research report that SUS is bi-dimensional having two components: usability (items 1, 2, 3, 5, 6, 7, 8, and 9), and learnability (items 4 and 10). This bi-dimensional nature of SUS has been found to be true in most of the testing scenarios [13, 41]. However, for the current case, the results are substantially different. Instead on loading separately on a distinct component items 4 and 10 loads together with other items (2, 3, 6, and 8). Items 1, 5, 7, and 9 load together on the second component; however, the loadings are low for items 7 and 9 (less than 0.5). Therefore, in case of the voice-assistants, the learnability component does not have any significance. The voice-only context provides a naturalistic and humanized environment when compared to the GUI systems that assists the users in completing their tasks. The low loading of item 7 ("I imagine that most people would learn to use this voice-assistant very quickly") further indicates that the learnability dimension is of little importance. In fact, for the SUS dataset when the factor analysis is re-run eliminating items 7 and 9, better results are obtained. Thus, for voice-assistants, SUS is reduced to an 8-item scale.

Finally, the correlation observed between the SUS and ARS scales is too low. ARS scale was built from SUS to give it an adjective rating and make the original SUS scores meaningful. Hence, historically SUS has shown to have a high correlation with ARS. Since ARS has just one item that measures the user-friendliness of a product, it is synonymous to the usability concept. However, the low correlation between the two scales is suggestive of the fact that SUS might not be a good measure of usability for the voice-assistants.

The VUS Scale

The factor analysis on the initial pool of items suggests three main components. These have been named as *Usability*, *Affective,* and *Recognizability & Visibility*. *Usability* dimension refers to the users' perceptions that the voice-assistants recognize them properly and do the tasks as instructed. This is related to the voice-assistant's ability to correctly recognize what the users are speaking, correctly interpret the

meaning of what the users are asking and act accordingly. In case of certain tasks that require a series of back and forth conversations between the user and the voice-assistant (for example during shopping or making a payment) for accomplishing the task, the users must know when to speak exactly. In the absence of such a scenario, there will be a lack of synchronization, which will make the system difficult to use. Therefore, this dimension is not only related to how easily the voice-assistants and the users understand each other, but also being able to interact freely and easily that makes these systems easy to use. This component accounts for the greatest proportion of the variance explained, indicating that it is one of the prime factors for evaluating the usability of voice-assistants. These days, voice-assistants are being used for a variety of purposes, both transactional and non-transactional. As such, the interactions must be clear, transparent and the information provided by these devices useful and timely.

The second dimension is *Affective*. We named it the affective dimension, since it portrays the satisfaction/frustration/expectation realization of the users after using the voice-assistants. This dimension explained the second-most proportion of variance in the factor analysis. The heuristic design principles for voice-assistants suggested by authors in [15, 46] also illustrate the importance of this factor. The anthropomorphic features of voice can easily cause a disconfirmation in the users' perceptions of the capabilities of a voice-assistant versus its actual abilities. Moreover, it has been found that voice-based systems are typically sluggish than GUI's [15], because the interaction is through voice-only. The naturalness and spontaneity of voice conversations increase the usability challenge for the voice-assistants as the users are accustomed to a certain way of communicating with other human beings and expect the same from these devices also. This makes the affective dimension highly relevant too for the purpose of usability evaluation.

The third and final factor is named as *Recognizability & Visibility*. Users must recognize the various functions and options provided by the voice-assistants just through interaction and affordance with the voice-assistants. The response given by the voice-assistants should be natural and easily understood by the users. In this respect, the choice of appropriate vocabulary is important that can be understood by a variety of users. Previous work in [37] also indicated problems with usability with respect to the accent of English spoken. Since, the voice-assistants lack any type of a visual interface, it can lead to a higher cognitive load among the users as they must remember all the speech commands. This might make these systems difficult to customize based on the needs and preferences of the users. The users may have to make many guesses while trying to customize the system using some trial-and-error method and might eventually abandon the task. Therefore, the visibility of the entire system might be affected.

Conclusion

Using standardized tools for usability measurement is an important aspect of any scientific and engineering process. Developing a standardized scale requires substantial efforts and a number of iterations; however, once complete, they are easy to reuse. The primary objective of this work was to check the suitability of SUS as a standardized measurement tool for voice-assistants, along with proposing a novel VUS scale that is targeted specifically for the voice-assistants. The current work reports on the first in a number of planned iterations in the development of the VUS scale until it is refined and matured. An EFA on the initial question pool suggests three main factors that contribute to the users' experience with the voice-based systems. We name these factors as *Usability*, *Affective,* and *Recognizability & Visibility*. Additionally, it is also found that the widely popular SUS may not be the best measure for evaluating the usability of voice-assistants, as its design principle is deeply rooted under a GUI environment. Overall, the results are promising, although there are scopes for further improvement.

However, it should be noted that this work is not without limitations. The limitations, in turn, pave the path for future work. The first limitation of the current work is using one voice-assistant only (Amazon *Alexa*). Although *Alexa* is the most popular voice-assistant being used currently, still including devices from other manufacturers could have introduced more variety. Therefore, all the current results are valid only for interactions with *Alexa*. Second, the sample size for the current study is 61. Although this number is in agreement with the Central Limit Theorem for getting statistically significant results, yet larger sample size should be able to capture more variations and establish more patterns. Third, in relation to usability assessment of voice-assistants, the number of published results is far from few. With respect to voice, currently, there is no published standard dataset with which comparisons can be made, which can be done for example in case of GUI's. Moreover, subjective experiments are difficult to be conducted and time consuming. On top of this, the users want to have a comparable experience while interacting with the voice-assistants, at par with their experiences of talking to another human-being. However, the voice-assistants have certain technical limitations, especially for understanding natural languages that makes it a challenge to design a suitable experiment. It will be helpful not only to compare the usability results from different voice-assistants to uncover newer usability themes, but also enable the system developers to evaluate the systems in isolation. However, for doing this, future research must focus on collecting

more normative data from a wider variety of applications as well as different types of users. Various user demographics like age, gender, experience, etc. might have an influence on the usability aspect. The current study just provides an initial attempt to show that the usability evaluation is different for GUI and voice-based systems, and that an iterative process of standardized scale development must be followed in a systematic manner for progressing on this aspect.

Funding This research is funded by the KMUTT New Researcher Funding

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Statista, "Smart Speaker Market Value Worldwide 2014–2025," [Online]. https://www.statista.com/statistics/1022823/worldwide-smart-speaker-market-revenue/, Accessed 8 Sep 2020.
2. Ki CW, Cho E, Lee JE. Can an intelligent assistant (IPA) be your friend? Para-friendship development mechanism between IPAs and their users. Comput Hum Behav. 2020;111:1–10. https://doi.org/10.1016/j.chb.2020.106412.
3. Statista, "Factors surrounding preference of voice assistants over websites and applications, worldwide," [Online]. https://www.statista.com/statistics/801980/worldwide-preference-voice-assistant-websites-app/, Accessed 8 Sep 2020.
4. McLean G, Frimpong KO. Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. Comput Hum Behav. 2019;99:28–37. https://doi.org/10.1016/j.chb.2019.05.009.
5. Pal D, Arpnikanondt C, Funilkul S, Chutimaskul W. The adoption analysis of voice based smart IoT products. IEEE Internet Things J. 2020;7(11):10852–67. https://doi.org/10.1109/JIOT.2020.2991791.
6. Feng L, Wei W. An empirical study on user experience evaluation and identification of critical UX issues. Sustainability. 2019;11(8):1–19. https://doi.org/10.3390/su11082432.
7. Oliver RL. A cognitive model of the antecedents and consequences of satisfaction decisions. J Mark Res. 1980;17(4):460–9. https://doi.org/10.2307/3150499.
8. Bhattacharjee A. Understanding information systems continuance: an expectation-confirmation model. MIS Q. 2001;25(3):351–70. https://doi.org/10.2307/3250921.
9. Parasuraman A, Zeithaml VA, Berry LL. A conceptual model of service quality and its implications for future research. J Mark. 1985;49(4):41–50. https://doi.org/10.2307/1251430.
10. Kocaballi AB, Laranjo L, Coiera E. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In: Proc. of 32nd international BCS human computer interaction conference (HCI) July 4–6; 2018. pp. 1–12.
11. Lewis JR. Standardized questionnaires for voice interaction design. ACIXD J. 2016;1(1):1–16.
12. Lewis JR. Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. Int J Hum-Comput Interact. 2018;35(15):1404–19. https://doi.org/10.1080/10447318.2018.1533152.
13. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact. 2018;34(7):577–90. https://doi.org/10.1080/10447318.2018.1455307.
14. Kocaballi AB, Laranjo L, Coiera E. Understanding and measuring user experience in conversational interfaces. Interact Comput. 2019;31(2):192–207. https://doi.org/10.1093/iwc/iwz015.
15. Murad C, Munteanu C, Cowan BR, Clark L. Revolution or evolution? Speech interaction and HCI design guidelines. IEEE Pervasive Comput. 2019;18(2):33–45. https://doi.org/10.1109/MPRV.2019.2906991.
16. Cowan BR, et al. What can i help you with?: Infrequent users' experiences of intelligent personal assistants. In: Proc. 19th international conference on human-computer interaction with mobile devices and services; 2017. pp. 1–12. https://doi.org/10.1145/3098279.3098539.
17. Silva AB, et al. Intelligent personal assistants: a systematic literature review. Expert Syst Appl. 2020;147:1–14. https://doi.org/10.1016/j.eswa.2020.113193.
18. Kawase T, Okamoto M, Fukutomi T, Takahashi Y. Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition. IEEE Trans Consum Electron. 2020;66(2):125–33. https://doi.org/10.1109/TCE.2020.2986003.
19. Kumar AJ, Schmidt C, Kohler J. A knowledge graph based speech interface for question answering systems. Speech Commun. 2017;92:1–12. https://doi.org/10.1016/j.specom.2017.05.001.
20. Guo L, Wang L, Dang J, Liu Z, Guan H. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. IEEE Access. 2019;7:75798–809. https://doi.org/10.1109/ACCESS.2019.2921390.
21. Nath RK, Bajpai R, Thapliyal H. IoT based indoor location detection system for smart home environment. In: 2018 IEEE international conference on consumer electronics (ICCE), Las Vegas, NV; 2018. pp. 1–3. https://doi.org/10.1109/ICCE.2018.8326225.
22. Greene S, Thapliyal H, Carpenter D. IoT-based fall detection for smart home environments. In: 2016 IEEE international symposium on nanoelectronic and information systems (iNIS), Gwalior; 2016. pp. 23–28. https://doi.org/10.1109/iNIS.2016.017.
23. Sun T. End-to-end speech emotion recognition with gender information. IEEE Access. 2020;8:152423–38. https://doi.org/10.1109/ACCESS.2020.3017462.
24. Park J, Son H, Lee J, Choi J. Driving assistant companion with voice interface using long short-term memory networks. IEEE Trans Ind Inform. 2019;15(1):582–90. https://doi.org/10.1109/TII.2018.2861739.
25. Jia J, et al. Inferring emotions from large-scale internet voice data. IEEE Trans Multimedia. 2019;21(7):1853–66. https://doi.org/10.1109/TMM.2018.2887016.
26. Alepis E, Patsakis C. Monkey says, monkey does: security and privacy on voice assistants. IEEE Access. 2017;5:17841–51. https://doi.org/10.1109/ACCESS.2017.2747626.
27. Zhang R, Chen X, Wen S, Zheng X, Ding Y. Using AI to attack VA: a stealthy spyware against voice assistances in smart phones. IEEE Access. 2019;7:153542–54. https://doi.org/10.1109/ACCESS.2019.2945791.
28. Malik KM, Javed A, Malik H, Irtaza A. A light-weight replay detection framework for voice controlled IoT devices. IEEE J Sel Top Signal Process. 2020;14(5):982–96. https://doi.org/10.1109/JSTSP.2020.2999828.
29. Yan C, Zhang G, Ji X, Zhang T, Zhang T, Xu W. The feasibility of injecting inaudible voice commands to voice assistants. IEEE Trans Dependable Secure Comput. 2019. https://doi.org/10.1109/TDSC.2019.2906165.
30. Thapliyal H, Ratajczak N, Wendroth O, Labrado C. Amazon echo enabled IoT home security system for smart home environment,

2018. In: IEEE international symposium on smart electronic systems (iSES) (Formerly iNiS), Hyderabad, India; 2018. pp. 31–36. https://doi.org/10.1109/iSES.2018.00017.
31. Nguyen QN, Ta A, Prybutok V. An integrated model of voice-user interface continuance intention: the gender effect. Int J Hum-Comput Interact. 2019;35(15):1362–77. https://doi.org/10.1080/10447318.2018.1525023.
32. Yang H, Lee H. Understanding user behavior of virtual personal assistant devices. IseB. 2019;17:65–87. https://doi.org/10.1007/s10257-018-0375-1.
33. Pal D, Arpnikanondt C, Funilkul S, Razzaque MA. Analyzing the adoption and diffusion of voice-enabled smart-home systems: empirical evidence from Thailand. Univers Access Inf Soc. 2020. https://doi.org/10.1007/s10209-020-00754-3.
34. Maguire, M. Development of a heuristic evaluation tool for voice user interfaces. In: Proc. of international conference on human-computer interaction (HCII'19), Orlando, USA; 2019. pp. 212–25. https://doi.org/10.1007/978-3-030-23535-2_16.
35. López G, Quesada L, Guerrero LA. Alexa vs. Siri vs. Cortana vs. Google assistant: a comparison of speech-based natural user interfaces. In: Proc. of 2017 international conference on applied human factors and ergonomics (AHFE 2017), Los Angeles, USA; 2017. pp. 241–50. https://doi.org/10.1007/978-3-319-60366-7_23.
36. Bogers T, et al. A study of usage and usability of intelligent personal assistants in Denmark. In: Proc. of international conference on information in contemporary society (iConference 19), Washington, USA; 2019. pp. 79–90. https://doi.org/10.1007/978-3-030-15742-5_7.
37. Pal D, Arpnikanondt C, Funilkul S, Varadarajan V. User experience with smart voice assistants: the accent perspective. In: Proc. of 2019 10th international conference on computing, communication and networking technologies (ICCCNT), Kanpur, India; 2019, pp. 1–6. https://doi.org/10.1109/ICCCNT45670.2019.8944754.
38. Ghosh D, Foong PS, Zhang S, Zhao S. Assessing the utility of the system usability scale for evaluating voice-based user interfaces. In: Proc. of the sixth international symposium of Chinese CHI (ChineseCHI '18), association for computing machinery, New York, NY, USA. pp. 11–15. https://doi.org/10.1145/3202667.3204844.
39. Yang C, Chen X, Xu Q, Hu L, Jin C, Wang C. A questionnaire for subjective evaluation of the intelligent speech system. In: Proc. of 2018 first Asian conference on affective computing and intelligent interaction (ACII Asia), Beijing; 2018. pp. 1–6. https://doi.org/10.1109/ACIIAsia.2018.8470339.
40. Brooke J. SUS: a quick and dirty usability scale. 1st ed. London: Taylor & Francis; 1996.
41. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum-Comput Interact. 2008;24(6):574–94. https://doi.org/10.1080/10447310802205776.
42. Ghosh D, Foong PS, Zhao S, Chen D, Fjeld M. EDITalk: Towards designing eyes-free interactions for mobile word processing. In: Proc. of the 2018 CHI conference on human factors in computing systems—CHI'18, association for computing machinery, New York, NY, USA. pp. 1–10. https://doi.org/10.1145/3173574.3173977.
43. Babel M, McGuire G, King J. Towards a more nuanced view of vocal attractiveness. PLoS ONE. 2014;9(2):e88616. https://doi.org/10.1371/journal.pone.0088616.
44. Koreman J, Pützer M. The usability of perceptual ratings of voice quality. In: Proc. 6th international conference on advances in quantitative laryngology, voice and speech research (AQL), Hamburg, Germany; 2003.
45. Pfeuffer N, Benlian A, Gimpel H, Hinz O. Anthropomorphic information systems. Bus Inf Sys Eng. 2019;61:523–33. https://doi.org/10.1007/s12599-019-00599-y.
46. Wei Z, Landay JA. Evaluating speech-based smart devices using new usability heuristics. IEEE Pervasive Comput. 2018;17(2):84–96. https://doi.org/10.1109/MPRV.2018.022511249.
47. Bangor A, Kortum P, Miller J. Determining what Individual SUS scores mean: adding an adjective rating scale. J Usability Stud. 2009;4(3):114–23.
48. Field A. Discovering statistics using IBM SPSS statistics. 4th ed. Thousand Oaks: Sage Publications; 2013.