IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 January 2024, accepted 19 January 2024, date of publication 25 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358423

SURVEY

A Review of Subjective Scales Measuring the User Experience of Voice Assistants

LAWAL IBRAHIM DUTSINMA FARUK[1], MOHAMMAD DAWOOD BABAKERKHELL[2], PORNCHAI MONGKOLNAM[1], VITHIDA CHONGSUPHAJAISIDDHI[1], SUREE FUNILKUL[1], AND DEBAJYOTI PAL[3]
[1]School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
[2]Department of Information Technology, Shaikh Zayed University (Khost), Khost 2501, Afghanistan
[3]Innovative Cognitive Computing Research Center (IC2), School of Information Technology, King Mongkut's University of Technology, Bangkok 10140, Thailand

Corresponding authors: Mohammad Dawood Babakerkhell (dawood.csf@gmail.com) and Debajyoti Pal (debajyoti.pal@gmail.com)

This work was supported by the Petchra Pra Jom Klao Doctoral Scholarship Research Funding through the King Mongkut's University of Technology Thonburi (KMUTT).

ABSTRACT The use of Voice Assistants (VA) in both commercial and personal contexts has experienced significant growth, emphasizing the importance of assessing their user experience (UX) for long-term viability. Currently, the development of appropriate scales that capture user viewpoints after interacting with a system has become a popular method for measuring UX of the Graphical User Interface (GUI) systems. However, the applicability of these scales that are meant for GUI systems on VA is still questionable, hence the need for analyzing the nature of previous scales used for measuring UX of VA. Additionally, in order to keep track of the state of UX research in the VA domain, it is crucial to understand the dimensions of UX that are being utilized. In this study, a comprehensive Systematic Literature Review (SLR) was carried out to identify 21 individual scales used for measuring UX of VA. Furthermore, this study present the evaluation criteria for assessing the rigor of operationalization during the development of these scales. The study analysis reveals that the scales used for measuring UX of VA extends beyond the traditional VUDA (value, usability, desirability, adoptability) principles and incorporates novel aspects such as anthropomorphism and machine personality. Future VA UX researchers should also acknowledge the variations in the rigorous measures employed during scale development, notwithstanding some common and accepted practices. Consequently, an overview is provided, along with suggestions for prospective studies in the field of VA UX research.

INDEX TERMS Factor analysis, reliability, scale, user experience, voice assistant.

I. INTRODUCTION

In the latter part of 1985, a well-received science fiction film titled 'Back to the Future II' was released [1], introducing the general public to the potential and possibilities of voice assistants. Today, it is an undeniable fact that what was once a technological fantasy has become a reality in people's homes. The public has witnessed the remarkable transformation of once-imaginary features into real existence. Voice assistants (VA), also known as intelligent personal assistants, are computer programs designed to understand human inquiries and engage in conversations with users using natural human language [2]. VA stand out as a unique and innovative tool compared to previous technologies because they don't rely on the traditional graphical user interface (GUI); instead, they use voice as their primary mode of interaction. The growing popularity of VA in personal spaces can be attributed to their seamless integration into people's daily routines. They handle fundamental tasks such as playing music, making dinner reservations, and setting reminders with ease [3]. Moreover, they are now used for more advanced technological activities like autonomous driving and tasks that require intensive cognition [4]. Various sectors, including the blind community [5], the educational sector [6],

The associate editor coordinating the review of this manuscript and approving it for publication was Giacinto Barresi.

customer and retail services [7], and the health sector [8], are harnessing the benefits of VA. The flexibility and efficacy of VA design make them highly desirable and affirm their limitless future potential. According to a report by [9], the projected market value of VA designed exclusively for personal use is a remarkable 8.4 billion US dollars by 2024. However, despite the initial enthusiasm surrounding VA, their unique and novel mode of interaction (voice) has sparked numerous inquiries and conjectures concerning their user experience (UX) [10]. Therefore, it has become crucial to study and investigate the UX of VA, given their significant relevance in the realm of interactive technologies and the world at large [11].

A. USER EXPERIENCE OF VOICE ASSISTANTS

The term "user experience" (UX) has recently gained more frequent usage and popularity. User Experience (UX) stands as a pivotal concept in Human-Computer Interaction (HCI), with the scientific community proposing various definitions for it [12]. However, the ISO standard definition is " UX is a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" [13]. On that accord, Individuals, academics, and professionals alike often employ the term "UX" to encompass all aspects of a user's experiences that occur before, during, and after interacting with a product, system or service. Essentially, UX also includes technical elements, value, usability, expectations, and satisfaction [11]. While determining what constitutes as a good UX remains subjective, studies have pointed out a positive UX is achieved when a product's design effectively aligns with the user needs [14]. Nevertheless, before delving into the realm of UX, it is important the reader briefly understand the concept of "user acceptance" because they are often misconstrued. User acceptance has been a well-established notion dating back decades ago, with well-defined models such as the Technology Acceptance Model (TAM) [15] and the Unified Theory of Acceptance and Use of Technology (UTUAT) [16]. However, while user acceptance and UX are closely related, they still remain distinct concepts. User acceptance represents the readiness and inclination of individuals to embrace and use a specific product, system, or technology. In contrast, UX centers on the holistic encounter of using, or anticipation of using a product or system. Nonetheless, a favorable UX plays a pivotal role in fostering heightened user acceptance [17].

The concept of UX can be assessed through three approaches; the measurement approach, the empathic approach, and the pragmatic approach [19]. Hence this review study focuses on the UX measurement approach, which is widely employed and effectively conducted using questionnaire scales [20].

The significant distinction between Graphical User Interfaces (GUI) and voice user interfaces (VUI) has led to a lack of consensus within the research community, regarding the applicability of scales specifically designed for measuring GUI UX use on voice user interface systems such as VA [21], [22]. Nevertheless, on that note, it is important to point out that some studies still utilize scales that were not originally designed for voice user interfaces to measure the UX connected to voice assistants [23], [24], [25]. Therefore, there is a growing interest in developing scales that are specifically tailored for voice user interfaces. However, at present, only a limited number of studies focused on assessing, and analyzing the properties and attributes of the existing scales used for measuring UX of VA [19], and none focus on their development. This suggests there is still a wealth of knowledge to be uncovered and explored in the VA UX domain. In light of this, the current review study is conducted to comprehensively examine the properties and development methods of scales currently use for assessing the UX of VA. The intention is to provide a diverse and thorough understanding of the topic, which can be valuable for making well-informed decisions.

B. SCALE DEVELOPEMENT

Scales are frequently utilized tools in the UX measurement approach. The use of scales in assessing latent constructs ensures precise research outcomes and delivers accurate measurements while mitigating bias [26]. Moreover, scales evaluate behavioral, attitudinal, and hypothetical aspects that may arise from individuals mental models, which cannot be directly observed) [27]. Recently, there has been a growing enthusiasm for development of UX scales to measure emerging technologies, such as chatbot [28], virtual reality [29], haptics [30], game design [31], augmented reality [58], and VA [32]. Nevertheless, even though VA is considered as a part of the emergent technology, there still remain a knowledge gap in understanding their UX scale development. Scale development is a rigorous process that encompasses many essential and yet intricate procedures. When executed poorly it can result in time-consuming efforts, significant economic expenses, and a vulnerability to inaccurate statistical analyses. Consequently, which ultimately jeopardize the overall validity and reliability of the scale [33]. Apart from the apparent drawbacks such as time consumption, other notable issues associated with poor scale development include; bias [34], item generation-related errors, and dimensions or concepts identification issues. Moreover, even everything goes well, choosing the proper method of analysis could be challenging [33], [35]. Therefore, to combat these issues, it is essential to understand the criteria employed while developing successful existing scales. This will aid in learning and understanding aspects such as the scale dimensionality, the techniques predominantly used, and the state-of-the-art analysis employed. Additionally, this will also prevent upcoming researchers and practitioners from falling into pitfalls of scale development problems.

Through answering the four research questions outlined below, this study aim to highlight what has been carried out in the realm of developing scales currently used for measuring the UX of voice assistants, and also provide valuable insights and guidance for future researchers and practitioners creating effective UX scales that can be used on VA.

RQ$_1$: What is the background, scope, and properties of studies that developed existing scales used for measuring UX of VA?

RQ$_2$: What conceptual dimensions are currently present on scales used for measuring UX of VA?

RQ$_3$: What methodological and operationalization rigor were employed while developing scales used for measuring UX of VA?

RQ$_4$: What are the reported reliability and validity of the current scales used for measuring UX of VA?

As evident from the research questions, this study intend to acquire an in-depth comprehension of the scale dimensions, attributes, developmental procedures, and the specific aspects of UX of existing scales used for measuring UX of VA. Additionally, every work included in this study will be distinctive, unique, and will contribute a diverse set of information, which collectively will enhance the quality of the review.

This study has been structured as follows: Section II presents the literature review carried out, which justifies the novelty of the study. The methodology employed in the study is outlined in Section III. Section IV presents the overall assessment criteria (framework) used for indicating what has been carried out during the development of the scales. The study result is presented in section V, followed by an in-depth discussion in section VI. Finally, section VII concludes the study, contain the study limitation, and provides recommendations for future work.

II. LITERATURE REVIEW

This section is dedicated to providing an overview of the current state of UX in Human-Computer Interaction (HCI) in general, and also reviewing previous review studies that has been carried out concerning conversational agents UX. While the main focus of this review study is an analysis of properties and development practices of scales used for measuring UX of VA, this section also briefly highlight some existing research on UX in emergent technologies.

A. USER EXPERIENCE OF EMERGENT TECHNOLOGIES

User experience measurement in the field of HCI is crucial in today's rapidly evolving technology landscape [36]. It serves as a critical determinant of the success and widespread adoption of innovative technologies, including voice assistants [37], virtual reality (VR) [38], chatbot [39], gesture-based communication systems [40], and more. A seamless and enjoyable UX can significantly lower entry barriers for new technologies, improving accessibility and user acceptance. Studies have extensively explored various aspects of UX in different emerging technologies. For instance a study by [40] explored the UX in gesture-based communication, aiming to enhance the quality of participants' daily lives through the use of an arm-swing input device. The study revealed that some individuals prefer gestures for communication because it's easy to learn and they enhance daily life interactions. Additionally, some studies have compared the UX of using different mode of interactions, which result in diverse conclusions. For instance [38] indicated there is significantly different experiences when using different interaction technologies. However, [41] conducted a comparison analysis between gesture-based interaction to touch-based interaction when controlling in-vehicle information systems (IVIS). Overall, the study found that gesture-based interaction can be a viable alternative to touch-based interaction, particularly in terms of trust and effectiveness when controlling in-vehicle information systems. Another area of UX gaining recognition is the focus on user group, such as integrating culture and tradition into emergent technology. A study by [42] investigated the influence of integrating traditional cultural elements into interaction design with the aim of enhancing UX. The study demonstrated that incorporating cultural elements into interaction design can indeed enhance the UX, improving interface usability and task completion rates.

Voice assistants, much like other emerging technologies, rely on UX design to ensure natural, efficient, and gratifying interactions [43]. Their ability to engage in human-like conversations enhances their practicality and appeal, yet this aspect is still not fully understood. The need for a review study is to consolidate existing knowledge across various dimensions of scales used for measuring UX of VA. Some review studies have already explored specific aspects of voice assistant UX. However, before embarking on such a study, it is crucial to ascertain that no previous review study has been carried out that cover our study objectives. This consideration leads us to the subsequent section.

B. PREVIOUS REVIEW STUDIES ON UX OF VA

Table 1 presents a collection of review studies conducted on conversational agents concerning UX, each review study cover a unique aspect of UX. Even though this study focus on scales used for measuring UX of VA, we also incorporate review study conducted on chatbot. Textual chatbot and voice assistants are both forms of conversational agents that utilize natural language processing and artificial intelligence to engage with users. By scrutinizing previous literature review studies on both voice assistants and chatbot user experiences alike, we will have holistic understanding of findings or design principles that can be applicable for both forms. Additionally, literature review studies such as [44] and [45] explores the use of both forms of agents. Finally, by including review studies on chatbots, it may reveal whether studies similar to ours have been conducted in the chatbot domain. To summarize the previous review studies carried out, Some reviews focus on specific user groups, such as older adults [46], while others are more broadly

TABLE 1. Summary of the previous review studies.

Study	Overview	Technology	Usage Scenario	Conclusion
[51]	This study highlights the challenges and limitations of adopting chatbot for commercial use.	Chatbot	Commercial.	In spite of the advancement in technology, the AI chatbot is still unable to stimulate human speech, due to poor dialogue modeling and limited domain-specific data.
[44]	This study analyzes the effectiveness and usability of conversational agents used in healthcare services.	VA & Chatbot	Health care.	There are mostly positive and mixed feelings about conversational agents and chatbot used in healthcare. There is a high level of usability and satisfaction. However, there is a mixed feeling when it comes to their effectiveness.
[52]	This study analyzes the extent to which VA are employed in the older adults' community and measures their technology readiness level.	VA	Older adults.	Older adults mostly use personal VA for setting up reminders, searching for information, and checking the weather. The technology readiness level in the aging community is high.
[45]	This study aims to identify the design, framework, and tools used to analyze the usability of health conversation agents.	VA & Chatbot	Health care.	There is a strong relationship between usability and UX. Usability is associated with more technical aspects. UX addresses more subjective aspects, such as user satisfaction.
[53]	This study analyses the use of VA in tertiary education and their impact on learning.	VA	Education.	VA help students with time management, and access to valuable information. However, it is still a new aspect with a need of further research in order to improve the motivation and engagement level towards learning.
[28]	This study carries out a systematic literature review to define attributes that analyze the interaction quality of chatbot, and also it design, and developed a new standardize scale ChatBot Usability Scale (BUS) for measuring chatbot satisfaction.	Chatbot	Commercial.	There is a new list of attributes specifically developed to measure the satisfaction level of a chatbot, as well being used as a primary tool for assessment.
[54]	This study analyses previous studies on conversational agents by exploring the existing frameworks and evaluation methods.	VA & Chatbot	General purpose.	There are different advanced public frameworks used for building a VA and chatbot that do not require programming language. These tools are able to detect intents and entities swiftly and provide answers using hidden heuristics.
[55]	This study carries out a review to analyze studies focusing on conceptualizing and operationalizing anthropomorphism, its antecedents, and consequences.	VA & Chatbot	General purpose.	Anthropomorphism plays a positive role in shaping the perception and adoption intention of these agents. Limited studies have explained how and why anthropomorphism exerts insignificant or negative effects, without highlighting whether there is an improper combination of anthropomorphic features.
[56]	This study analyzes the current literature on COVID-19-related chatbot to identify, characterize, and highlight their challenges.	Chatbot	Health care.	Using a chatbot to fight COVID-19 is still in its early stages, and researchers need to understand the domain knowledge better to make the chatbot effective and useful to fight Covid-19.
[46]	This study identifies the state-of-the-art experimental studies with a chatbot with a screen display capable of verbal dialogues, focusing on older adults with amnesia.	Chatbot	Older adults.	There are still limited studies that focus on chatbot that support people with memory problems.
[47]	This study highlights the dimensions, the independent variables, and the environment employed by studies while measuring the usability of VA.	VA	General purpose.	There are diverse dimensions used for measuring the usability of VA, with some being employed more frequently than others. The ISO 9241-11 framework is not suitable to measure the current usability dimensions due to the advancement and more user expectations.
[50]	This study conducts a meta-synthesis on the conversational agent's voice both from the design and experience aspect, based on a human-centered perspective.	VA	General purpose.	Understanding how people engage and are also influenced by the machine voice is important. The voice of the machine is perceived, contextual, and dynamic which has a huge impact on the voice-based human-agent interaction.
[48]	This study presents a review to investigate how the UX is assessed when interacting with VA.	VA	General purpose.	Majority of the studies used their unique developed evaluation methods, without using any questionnaires validated for UX evaluation. Some studies used evaluation tools before participants interacted with agents, and only a minor handful carried out assessments prior, during, and post use.
[49]	This study carries out a rapid review of previous works to analyze factors that are manipulated as independent variables during VA interactions and the usability dimensions.	VA	General purpose.	There are dimensions such as affective, trust, and sociality which play a major role in the interaction with VA. There is a strong dependency on subjective methods of evaluation. Objective methods need to be researched more.

oriented towards general usage [47]. As the field of VA is still relatively new, most of these review studies that were carried out on UX of VA have focused on the applicability of VA in a specific areas like education and healthcare [44]. However, only a small number of review studies (n=2) [48], [49] have taken a broader view of examining aspects such as the dimensions of the scales used for measuring UX of VA, their frameworks and independent variables. Nevertheless, similar to other review studies in this field, they did not focus on the scale development aspect.

Similarly, a subset of studies (n=4) specifically focuses on the usability aspect of VA [44], [45], [47], [50] which is just one facet of UX. Consequently, the treatment of the UX analysis still remains somewhat nonexistent or superficial. Hence, this study represents the inaugural effort to systematically review, and conceptually analyze the attributes and development practices of scales used for measuring UX of VA.

III. METHODOLOGY

A review study approach was selected in order to analyze, evaluate, and interpret the pertinent articles that will address the identified research questions. Therefore, the collection of relevant articles is crucial, for it entails a meticulous and diligent process to ensure the inclusion of only relevant articles. The methodology section outlines the various steps undertaken before selecting the resources for this study, as elaborated upon later in this section. However, before we commence our search query, it is important to clarify that the scales we intend to incorporate in our study can be either one of the three aspects. Firstly, scales specifically designed for voice assistants or voice user interfaces. Secondly, scales not originally developed for voice assistants or voice user interfaces but are utilized by studies to assess the UX of voice assistants due to their reliability, albeit with the caveat that they may not encompass all aspects of voice-based interaction. Lastly, scales not specifically created for voice assistants, yet suggested by scale authors for potential application in future studies involving voice assistants. We believe this will encompass a suitable range of scales that are employed for evaluating the UX of voice assistants.

A. SEARCH QUERY

Generating effective keywords and search queries is a crucial initial requirement for any review methodology. These carefully selected keywords ensure the inclusion of relevant and precise article that have previously developed scales used for measuring UX of VA as their outcome. Therefore, thorough examination preceded the choice of these keywords. To ensure the search query are more specific and inclusive of a wide range of studies, Boolean operators such as "AND" and "OR" were used. Moreover, wild characters (∗) were also used to include words that might have suffixes such as "measure*" which could represent either measure, measures, measurement. When selecting the appropriate keywords for the search, the researchers focused on three main aspects: the system which is VA in our case, the UX aspect, and scale development aspect. As highlighted earlier, the scope of this study revolves around scales used to measure voice assistants. Therefore, the term "Voice Assist*" was employed, the wild character * was included because to cover the different variations (assistant, assistants, assistance). VA is often referred to as intelligent assistants, personal assistants, or conversational agents; therefore, all these synonyms were considered in the search. Moreover, terms such as 'chatbot' were avoided since they primarily indicate a textual program. To encompass systems that use human-like simulated speech as the mode of communication within their scope, the term 'speech interface' was included. Regarding 'scale development', the term was selected as it pertains to the process of creating scales, and other relatable terms such as "measures" and "questionnaires" were also employed.

The Final outcome query of the search that was mutually agreed upon by all the authors was ("Voice Assist*" OR "Conversational Agent*" OR "Intelligent Personal Assistant*" OR "Speech Interface") AND ("System Usability" OR "Usability" OR "UX" OR "User Experience") AND ("Scale Development" OR "Measure*" OR "Questionnaire" OR "Scale"). This keyword query will bring back studies that develop scales used to measure UX of VA.

B. ARTICLE IDENTIFICATION AND SELECTION

There are numerous studies that present guidelines for conducting a successful article identification and selection process while carrying out a review study [26], [57]. On that aim, this review study adopt the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology proposed by [58] due to the method maturity and popularity. The PRISMA method consists of four stages for final article selection, namely: Identification, screening, eligibility, and inclusion. Fig. 1 presents the graphical representation of this study entire methodology based on the PRISMA standard.

1) IDENTIFICATION

Multidisciplinary (Scopus, Google Scholar) and individual publishers' databases (ACM, IEEE, Elsevier, Taylor and Francis, and Springer Link) were used for the article selection process. This study used more than the average number of database used in a systematic review [59]. The databases used contain relevant articles concerning scales used for measuring UX of VA. As stated earlier, UX of VA is a topic that is gaining fair amount of traction, which makes multidisciplinary database such as Scopus and Google scholar suitable as well. Likewise, publisher databases such as ACM, IEEE, Elsevier, Taylor and Francis, and Springer Link were also utilized, because they narrow down the search for relevant articles and increase the chances of retrieving full-text articles.

TABLE 2. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
The study should be on scales development that are utilized to measure the UX of VA or voice user interfaces.	Literature review studies
The study should be on scales created to measure UX or any aspect of it.	Studies that focus on testing, comparing, or evaluating existing scales.
The study can focus on the modification of previous scales.	Studies that include systems where voice is not a major component of the communication modality, furthermore the output scales have not been employed to measure the UX of voice assistant.
The study should be in English with an available full text.	

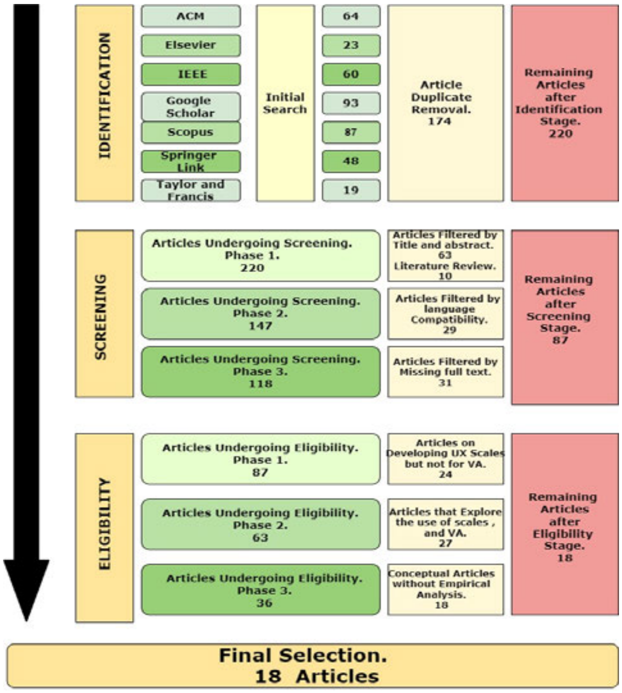
IDENTIFICATION
ACM
Elsevier
IEEE
Google Scholar
Scopus
Springer Link
Taylor and Francis
Initial Search
64
23
60
93
87
48
19
Article Duplicate Removal. 174
Remaining Articles after Identification Stage. 220
SCREENING
Articles Undergoing Screening. Phase 1. 220
Articles Filtered by Title and abstract. 63 Literature Review. 10
Articles Undergoing Screening. Phase 2. 147
Articles Filtered by language Compatibility. 29
Articles Undergoing Screening. Phase 3. 118
Articles Filtered by Missing full text. 31
Remaining Articles after Screening Stage. 87
ELIGIBILITY
Articles Undergoing Eligibility. Phase 1. 87
Articles on Developing UX Scales but not for VA. 24
Articles Undergoing Eligibility. Phase 2. 63
Articles that Explore the use of scales , and VA. 27
Articles Undergoing Eligibility. Phase 3. 36
Conceptual Articles without Empirical Analysis. 18
Remaining Articles after Eligibility Stage. 18
Final Selection. 18 Articles

FIGURE 1. Article selection process.

Moreover, these databases are highly regarded in the HCI community. In conclusion the search query returned 64 articles from the ACM database, 60 articles from the IEEE database, 48 from the Springer Link database, 23 articles from Elsevier database, 19 articles from Taylor and Francis database, 93 articles from Google Scholar database, and 87 from the Scopus database (394 articles in total). About 174 duplicate articles were removed, and 220 articles were moved to the screening stage.

2) SCREENING

The screening process was carried out to eliminate studies that did not meet the inclusion and exclusion criteria. The screening began with a thorough examination of the title and abstract of each study to identify prominent mismatches, leading to the removal of 63 articles for this reason. Additionally, 10 articles were excluded because they were literature review studies. About 29 articles were not written in English and 31 articles that were missing the full text were removed. As a result, 87 articles were selected for full-text reading to assess their eligibility. Table 2 provides an overview of the inclusion and exclusion criteria utilized in this process.

3) ELIGIBILITY

The eligibility procedure was carried out in a more in-depth and rigorous process than in the previous stage. The authors read the full texts of the 87 articles and remove the ones that did not fit into the study objectives. For instance, 24 articles were removed because they were on scale development, but not either on VA and voice was not a major component in the interaction [24], [60]. Furthermore, 27 articles were removed because they analyzed and explored the use of existing scales. For instance, [26] focus on evaluating the validity of the User Experience Questionnaire (UEQ) scale, rather than developing any modified version of UEQ. Finally, 18 articles were removed that did not provide any empirical analysis.

4) FINAL SELECTION

At the end of the article identification and selection procedure, 18 articles were selected for the study, which is roughly about 6.3% of the articles from the initial search, which indicating many false positives hits in the search process. The reasons for the occurrence of false positive matches during the search process may include, but not limited to the following factors: Firstly, VA UX is gaining popularity worldwide, leading to an increased demand for scales in multiple languages. Researchers are increasingly translating the original English versions into different languages to accommodate this need. Additionally, given the conversational AI nature of VA, interactions in languages other than English are required. Therefore, approximately 29 studies had their full texts in a non-English format, necessitating their exclusion. Secondly, the terms "conversational agent" or "intelligent personal assistant" are frequently used interchangeably by researchers, encompassing both voice and non-voice contexts (e.g., textual chatbot). However, this study focuses on the measuring voice communication modality as a core component of interaction, non-voice communication systems were excluded. As a result, approximately 97 articles were eliminated based on this criterion. Thirdly, the domain of UX with VA is relatively new. Several exploratory studies have been conducted to investigate human perceptions of existing scales, but empirical analyses have not been the primary focus.

It should be noted that, given the widespread popularity of some scales such as the System Usability Scale (SUS), Attrakdiff, Mean Opinion Score (MOS) and User Experience Questionnaire (UEQ); which were not originally created for voice assistant or voice user interface. However, HCI researchers tend to use these scales without modifications for voice-based system [24], [25], [61], [62]. Therefore they were added to the study.

IV. FORMATION AND EVALUATION OF CODING CRITERIA

In this section, the evaluation and coding criteria were formulated and presented. These criteria will serve as the analytical foundation for assessing the operationalization rigor employed during the development of the existing scales used for measuring UX of VA. The process and criteria employed in scale development are subjective and depends on the intent of the developer. However, they are widely accepted criteria in scale development. It is important to emphasize that this assessment criteria do not rank the scales quality or importance, neither do they claim one scale is better than the other. Instead, they are applied to systematically review and evaluate the process the authors used while creating their scales, as well as the operationalization rigor. Table 3 outlines the criteria used in this study. While each criteria outlined is crucial during the scale development process, researchers often do not carry out all of them. The evaluation and coding criteria were based on [26] and [60]. Normally, for each criterion, a score of 1 point is awarded if it is fully met, 0 points if there is no fulfillment, and in some cases 0.5 points if there is a partial fulfillment. Brief explanation for each criterion is given, highlighting its importance in the overall scale development process.

A. CONCEPTUAL DIMENSION

Conceptualizing the dimensions of the scales is an integral initial step in scale development. This is because dimensions act as representations of the constructs being targeted by the scale. A well-defined dimension should capture the fundamental concept of a construct at face value [27]. Moreover, it also delineates the focus of the scales [63]. However, there are instances where dimensions within the same scale or across different scales might be overly similar, leading to homonyms or synonyms among multiple dimensions [27]. Therefore, it becomes necessary to provide clear definitions for the dimensions in order to prevent any issues arising from ambiguity regarding their representation [60]. A point is assigned to studies that report their conceptual dimensions and an additional 1 point if a proper definition is provided for each dimension. Studies that do not report these processes receive no points.

B. ITEM GENERATION

Item generation, also known as question development, comes after conceptual dimension. Items should be brief yet comprehensive and must be a valid expression of the dimension they represent [64]. There are two methods commonly used for generating items: deductive and inductive methods. Researchers can choose to employ either method; however, it is considered the best practice to utilize both [26]. The inductive method involves qualitative techniques such as conducting interviews with focus groups or experts [27]. The deductive method, on the other hand, relies on literature reviews and the evaluation of existing scales. The inductive methods are not frequently employed, and it is uncommon for researchers to apply both the methods simultaneously [26]. A point is assigned to studies that report the use of deductive methods during item generation, and an additional 1 point to studies that report the use of inductive methods. Studies that fail to report the use of either method receive 0 points.

C. ITEM REFINEMENT

The item refinement stage is when an arbitrator examines the dimensions and items before they are presented to the participants. During the item generation process, various issues may arise, such as complex wording, redundancies, biased questions, vagueness, and questions requiring estimation [64]. Therefore, an impartial review of the dimensions and items by an arbitrator is necessary after successful item generation. The methods employed during item refinement can vary; however, the most commonly used methods include pretests, expert evaluations, and pilot tests. Pretesting is conducted on small samples to gather user feedback on the items. Pretesting can help address issues like ambiguity, complexity, word sensitivity, and missing items [60]. Furthermore, pretesting can involve focus groups, consisting of individuals, preferably from the target population. This typically involves conducting an open discussion with the researcher, including recording the process or taking notes. Another criterion for item refinement is expert evaluation. Unlike focus groups, experts are participants with domain knowledge or specialists in the subject matter. Expert feedback is essential for assessing the quality of the items and determining whether they accurately reflect the intended dimensions. Open-ended feedback is often sought when requesting expert evaluation [64]. Following the expert evaluation, a pilot test, also known as a pilot study, is typically conducted. A pilot test involves sending a small sample of the pretested and expert-evaluated items to a few participants, preferably a subset of the target population. The purpose of the pilot test is to evaluate how the items perform under actual field conditions for which they were developed.

All three stages of item refinement mentioned above are crucial; however, some researchers combine them into a

TABLE 3. The evaluation and coding criteria.

Evaluation Criteria	Coding	Assessment
Conceptual Dimension	Dimension name	1 point if reported; 0 point if not reported
	Dimension definition	1 point if reported; 0 point if not reported
Item Generation	Deductive method	1 point if reported; 0 point if not reported
	Inductive method	1 point if reported; 0 point if not reported
Item Refinement	Pretest	1 point if reported; 0 point if not reported
	Expert evaluation	1 point if reported; 0 point if not reported
	Pilot test	1 point if reported; 0 point if not reported
Sampling Accuracy	Sample size	1 point if the sample size is (n > 300) and the sample size to items ratio is (10:1); 0.5 point if the sample size is (n > 30); 0 points if none is satisfied
	Sampling technique	1 point if mentioned; 0.5 point if the sample technique is not specifically mentioned but hinted through stating the sample features; 0 point if not mentioned
Data Factorability	Correlation matrix	1 point if correlation matrix is presented; 0 point if study fails to report correlation matrix
	Bartlett's test of sphericity	1 point if the test of Bartlett's sphericity is reported with a probability of 0.5 or less; 0 point if Bartlett's test of sphericity is reported
	Kaiser-Meyer-Olkin (KMO) value	1 point if KMO is reported with values between 0.8 and 1; 0.5 point if KMO is carried out with values below 0.8; 0 point if KMO is not reported
Factor Analysis (Exploratory factor analysis)	Factor extraction	1 point if employed (any non-PCA based technique); 0.5 point if employed but PCA used for extraction purpose; 0 point if not employed
	Factor retention	1 point if reported; 0 point if not reported
	Factor rotation	1 point if reported; 0 point if not reported
Factor Analysis (Confirmatory factor analysis)	Dedicated samples	1 point if EFA and CFA carried out on separate samples; 0 if there was no highlight on the sample split
	Model fit indices	1 point if at least 3 fit indices are reported; 0.5 point if less than 3 fit indices are reported; 0 point if no fit indices are reported
Psychometric Properties (Validity)	Content Validity	1 point if reported; 0 point if not reported
	Construct Validity	1 point if convergent validity is reported; 1 point if discriminant validity is reported; 1 point if construct validity is reported without specifying the type; 0 point if not reported at all
	Criterion validity	1 point if predictive validity is reported; 1 point if concurrent validity are reported; 1 point if criterion validity is reported without specifying the form; 0 point if not reported at all .
Psychometric Properties (Reliability)	Internal Consistency	1 point if the internal consistency of each of the dimension is > 0.7; 0.5 point if the internal consistency is < 0.7; 0 point if not reported

single process [60]. A point is awarded when a pretest is conducted, another point when expert review is reported to have been carried out, and an additional point when a pilot test is reported to have been conducted. Studies that do not report any of these steps receive zero points.

D. SAMPLING ACCURACY

After preparing the dimensions and items, the evaluation of sampling accuracy begins. The quality of the sample plays a crucial role in factor analysis and should be carefully considered. A small sample size can lead to poor factor stability, resulting in biased scales [65], [66], and reduced generalizability [67]. Various studies have emphasized the importance of determining the appropriate sample size for scale development, but the ideal size remains subjective. While the determination of sample size is subjective and can be influenced by various external factors, this study established this criterion based on three commonly utilized sample size criteria: 1). the sample size must be at least 300 [68], 2). the sample size to items ratio must be at least 10:1 [69], and 3). the minimum sample size should be 30, as suggested by the Central Limit Theorem (CLT) [70].

One point was assigned to every study that reported a sample size meeting the first two criteria mentioned above. For studies that reported a sample size satisfying only the third criterion of at least 30, 0.5 points were awarded to them. Finally, studies that reported a sample size below the three sample criteria or did not report any information on the sample size received 0 points.

A sampling technique is employed to determine the selected samples. These techniques are typically classified as either probability or non-probability [71], each with its advantages and limitations. While this study does not endorse one sampling technique over the other, it assesses whether any sampling technique was reported by the study. Accordingly, 1 point is assigned when the sampling technique is explicitly mentioned, 0.5 points are assigned when the sampling technique is not explicitly reported but can be inferred from the sample characteristics. Finally, 0 points are awarded when the sampling techniques are neither reported nor inferred.

E. DATA FACTORABILITY

Following data collection, it is essential to evaluate the suitability of the collected data for factor analysis. This evaluation, known as data factorability, relies on commonly employed methods such as the correlation matrix, Bartlett's test of sphericity, and the Kaiser-Meyer-Olkin (KMO) measure [60]. The correlation matrix indicates the degree of linear relationship (correlation coefficient) between variables. Bartlett's test of sphericity is employed to determine if the correlation matrix is an identity matrix.

An identity correlation matrix suggests that the items are unrelated and do not represent a common underlying dimension, making them unsuitable for factor analysis. Bartlett's test of sphericity typically yields a significance level of 0.05 or less, indicating that the correlation matrix is not an identity matrix.

In contrast, the KMO measure assesses the correlation among items that share common underlying dimensions. KMO values approaching 1.0 are considered desirable, indicating a strong correlation between the items, while values below 0.5 are considered inadequate [26]. In this study assessment, studies that reported a correlation matrix analysis receive 1 point. Similarly, studies that report Bartlett's test of sphericity are awarded 1 point. Additionally, studies that report their KMO measure receive 1 point. On the other hand, studies that did not report any of these three methods are given 0 points.

F. FACTOR ANALYSIS

Factor analysis is not a single procedure but rather involves a variety of statistical and methodological choices. The selection of what method to utilize is subjective and depends on factors such as the developer's preferences, the features and representation of the data. There are two types of factors.

Analysis: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) [72]. EFA is a statistical method used to uncover the underlying structure of a relatively large set of variables. Studies employ EFA to identify the underlying relationships between the measured variables. Since EFA is a series of process, it was classified it into three principal procedures: factor extraction, factor retention, and factor rotation. The selection of the factor extraction method has a substantial influence on both the outcomes and interpretation of the factor analysis. Each factor extraction method, operates on distinct assumptions and methodologies, resulting in divergent outcomes in terms of the extracted factors [73]. One point each to studies that highlight factor extraction methods, e.g., maximum likelihood and principal axis factor. However, 0.5 points was awarded each to studies that employed Principal Component Analysis (PCA) as the factor extraction criterion since PCA is a technique more suitable for dimension reduction in a scale rather than uncovering latent factors [66]. One point was assigned to each study that explicitly mentioned or present the method used for factor retention, such as scree plot or parallel analysis. Furthermore, 1 point was awarded to study that conduct and mention the factor rotation procedures they employed, whether orthogonal or oblique. Confirmatory factor analysis (CFA) tests whether the developed model is consistent with prior hypotheses. The CFA occurs when an underlying hypothesized measurement structure exists between dimensions and the proposed items [74]. CFA is affected by two significant aspects: the CFA-dedicated samples and the model-fit indices. The sample collection procedure is still the same for the entire study; however, the sample used for CFA must be different from the sample used for EFA to validate the model. However, there is no standardized split ratio between the sample data. Therefore, 1 point was awarded when the authors mentioned that they have split the data for carrying out EFA and CFA and 0 points when they failed to mention anything, or no CFA has been conducted. The fit indices help validate and evaluate the model's fitness. The standard model fit indices reported in [75] was referenced on Table 4, which presents the commonly used and reported fit indices. One point was awarded to every study that reported at least three fit indices, 0.5 points each to studies that reported less than three fit indices, and no points to studies that did not report any fit indices.

G. PSYCHOMETRIC PROPERTIES (VALIDITY AND RELIABILITY)

Validity and reliability analysis are needed to be carried out to measure the legitimacy and consistency of a scale. Validity represents how accurately the scale's dimensions measure what they claim to measure [76]. This study focus on three main types of validity that are often used: Content, construct, and criterion. Construct validity measures how accurately the scales evaluate what it is meant to evaluate. This type of validity confirms if the scales align with the construct, they aim to measure [77]. Convergent and discriminant validity are the two types of construct validity, and both are necessary to provide evidence of valid measurement [78]. Content validity refers to the extent which a measurement instrument adequately covers the full range of items that should be included to represent a specific construct. It focuses on whether the items in a scale are relevant, comprehensive, and representative of the construct being measured.

Content validity is typically established through expert judgment and involves assessing the relevance and representativeness of the items in relation to the construct [26]. Lastly, criterion validity measures the concrete outcome [79]. The criterion validity can be measured using two approaches, predictive validity, and concurrent validity. Predictive validity measures how scales will perform when used in the future, and concurrent validity is used to measure the correlation between the new scales and well established existing scale [80].

Reliability refers to the degree to which the scale is consistent and stable in measuring what it is intended to measure. To put simply, scales are reliable if they are consistent within and across time. In general, there are four types of reliability: test-retest, parallel forms, inter-rater,

TABLE 4. Model-fit indices.

Model-Fit Criterion	Acceptable Level
Chi-square (χ^2)	Low $\chi 2$ relative to degrees of freedom with an insignificant p value ($p > 0.05$)
Goodness-of-fit index (GFI)	0 (poor) to 1 (fit); Values close to 0.90 or 0.95 considered as an appropriate fit.
Root-mean square residual (RMR)	Indicates the closeness to s matrices
Standardized RMR (SRMR)	Value less than 0.05 signify worthy model fit
Root-mean-square error of approximation (RMSEA)	Value of 0.05 to 0.08 signify worthy model fit
Tucker-Lewis Index (TLI)	0 (poor) to 1 (fit); Values close to 0.90 or 0.95 considered as an appropriate fit.
Normed fit index (NFI)	0 (poor) to 1 (fit); Values close to 0.90 or 0.95 considered as an appropriate fit.

and internal consistency. This study considers only internal consistency because it measures the consistency of individual items on a scale. The higher the internal consistency, the more reliable is the dimension. Researchers consider Cronbach's Alpha an appropriate measure for internal consistency. According to [81], there is a general concession on the acceptable value of Cronbach's Alpha. One point is assigned to studies that report content validity.

Furthermore, 1 point was awarded to studies that report the convergent form of construct validity and another 1 point to study that report the discriminant form of construct validity. Additionally, for criterion validity, 1 point was awarded to studies that report predictive form of criterion validity and 1 point to studies that report concurrent form of criterion validity. However, if a study did not report any criterion, nor construct validity, 0 points was awarded. Additionally, 1 point was awarded to every study that reported their dimension Cronbach's Alpha with values > 0.70, 0.5 points awarded to studies that reported their Cronbach's Alpha values but with some values < 0.70, and 0 points if no Cronbach alpha values were reported.

V. RESULT

In this section, the results of the data synthesis in accordance with the study objectives is presented. Consequently, the results are organized to align with the study research questions.

A. RQ$_1$: WHAT IS THE BACKGROUND, SCOPE, AND PROPERTIES OF STUDIES THAT DEVELOPED EXISTING SCALES USED FOR MEASURING UX OF VA?

Every article included in this study, developed a unique scale that measure some aspects of UX of voice interface system, and are utilized in the field of VA. Knowing the contextual properties of the previous studies will guide the design of future research on scales used for measuring UX of VA. Therefore, an overview of the properties of the included studies is presented in Table 5 and analyzed under the following heads: The type of voice interface used for experiment when developing the scales (tools), the geographical origin of the scale (population studied), what facet of UX the scale was developed to measure, the number of dimensions considered (a more in-depth analysis of the dimensions is carried out in RQ$_2$), the status of the scale (novel vs. iterative/modified version), and a highlights on studies that utilize the developed scales. This analysis will help understand the current state of the studies, also is essential for proper interpretation and application of the study's results. There were a total of 21 scales in total collected from 18 studies, because two different studies developed more than one scale as their output.

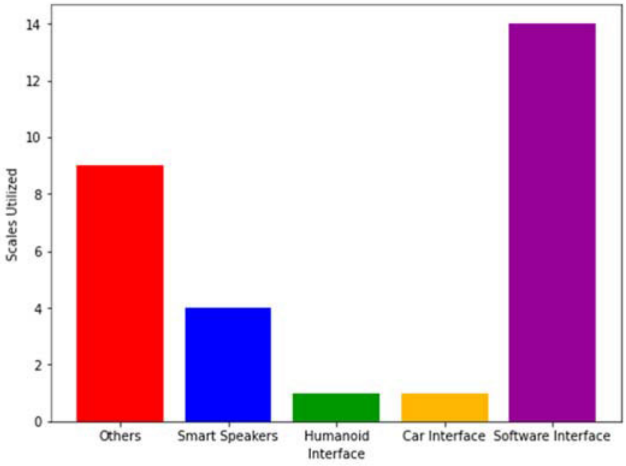

FIGURE 2. Different types of VA devices/interfaces.

1) TYPES OF DEVICE/INTERFACES

Various device/interfaces were used when developing scales because the choice of device interface induce different levels of mental cognition and affects the UX [109], [110]. However, the choice of devices might differ based on the technology available in that time period which the scale was created, and also the author's intention for developing the scale. Nevertheless, as evident from Fig. 2, multiple device interfaces are being utilized in the development of the current scales, with some scales incorporating multiple interface types. For instance, scales such as SASSI, UEQ-S,

TABLE 5. Summary of the current scales properties (no. of studies = 18, no. of scales = 21).

Scale Developed	Interface					Origin	Intended Measurement	No. of Dimensions	Status	Studies that Utilized Developed Scale.
	Smart Speaker	Humanoid	Car Interface	Software Interface	Others					
[82] SASSI			x	x	x	England	Usability	6	Novel	[23], [37], [83]
[84] Attrakdiff 1				x		Germany	Quality of use	3	Novel	[61]
[85] Attrakdiff 2				x		Germany	Quality of use	3	Modified	[23], [37]
[86] Mean Opinion Scale (MOS)				x		United States	Voice Quality	2	Novel	[62], [87]
[88] Mean Opinion Scale Expanded (MOS-X)				x		United States	Voice Quality	4	Modified	[88], [89]
Mean Opinion Scale-Revised (MOS-R)				x		United States	Voice Quality	2		[88]
Mean Opinion Scale Expansion (MOS-R3)				x		United States	Voice Quality	4		[88]
[90] Speech User Interface Service Quality (SUISQ)					x	United States	Usability Voice Quality Affective Response	4	Novel	[91]
[92] Speech User Interface Service Quality (SUISQ-R);					x	United States	Usability Voice Quality	4	Modified	[37], [93]
Maximally Reduced SUI Service Quality (SUISQ-MR)					x		Usability Voice Quality	4		[94]
[95] User Experience Questionnaire (UEQ)				x		Germany	Holistic experience	6	Novel	[25]
[96] User Experience Questionnaire (Short Version) UEQ-S				x	x	Germany	Holistic experience	2	Modified	[97]
[98] User Experience Questionnaire UEQ+					x	Germany	Holistic experience	3	Modified	[99]
[100] System Usability Scale (SUS)				x		United States	Usability	2	Modified	[23], [24], [37], [91], [93]
[101] User Experience Evaluation Method for Spoken and Multimodal Interaction (SUXES)	x			x	x	Finland	Holistic experience	9	Novel	[102]
[103] Conversational Agents Scale (CAS)				x		Germany	Anthropomorphism Holistic Experience	6	Novel	[103]
[104] Quality of Experience (QoE)	x	x		x	x	United Arab Emirates	Holistic Experience	8	Novel	[104]
[105] Conversational Agent's Usage Scale (CAUS)					x	Lithuania	User Interaction	3	Novel	[105]
[106] User Experience Evaluation of Conversational Systems (UEXECS)				x		Brazil	Holistic Experience Usability	10	Novel	[106]

TABLE 5. ***(Continued.)*** **Summary of the current scales properties (no. of studies = 18, no. of scales = 21).**

Scale Developed	Interface					Origin	Intended Measurement	No. of Dimensions	Status	Studies that Utilized Developed Scale.
	Smart Speaker	Humanoid	Car Interface	Software Interface	Others					
[32] Personality, Usability, and Enjoyability of Voice Agents (PUEVA)	x					United States	Usability Enjoyability Personality	3	Novel	[107]
[10] Voice Usability Scale (VUS)	x					Thailand	Usability	3	Novel	[108]

SUXES, and QoE are all scales developed using multiple interface types. It is noteworthy that car speech interfaces, which have gained attention recently [111], were employed in the development of one of the earliest scales to measure the UX of voice-based systems - the SASSI scale [82]. According to the result, only 9.5% of the scales (n = 2) were created with smart speakers, and 19% of the scales (n=4) were created using multiple interfaces. The majority, comprising 67% of the scales (n = 14), were created using diverse software interfaces, including text-to-speech [112], third party applications such as gaming applications, travel software, and customer relationship management software [92]. In terms of the publication year, 50% of the studies (n = 9) were published between 2000 and 2010, 11% of the studies (n = 2) were published between 2011 to 2015; and 39% of the studies (n = 7) were published between 2016 to 2022. The results suggest that developing scales for measuring voice interaction has been carried out since the past two decades. However, the early attempts were mainly targeted towards software systems like text-to-speech. It is only recently, in the past 5 years that researchers are trying to develop scales for emergent technology like smart speakers and humanoids.

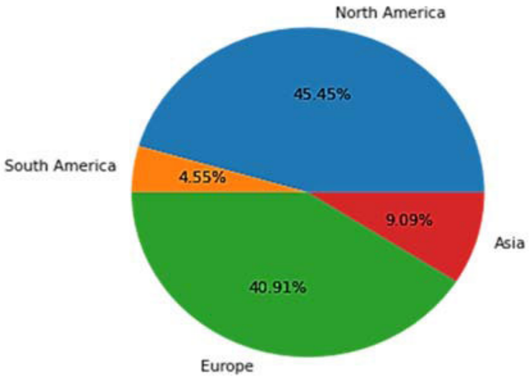

FIGURE 3. **Origin of the scales by continents.**

2) ORIGIN OF SCALE

The origin of the studies reveals a notable focus in majorly Europe and North America, with the target user population primarily being from United States of America and Germany. Specifically, approximately 43% of the scales (n=9) originated from the United States, 29% (n=6) from Germany, and 5% (n=1) each from Brazil, Lithuania, Thailand, England, Finland, and the United Arab Emirates. This continental variation is illustrated in Fig. 3. Surprisingly, none of the scales originated from Australia, despite the fact that around 34% of Australian customers prefer using VA to address inquiries than SMS or email [113]. As for Africa, VA adoption is still in its early stages due to technical challenges, speech patterns, and the use of non-accented English [114], which explain the absence of any scales from the region. As part of RQ_1, the study investigate the origin of the scales, considering the significance of culture and location in measuring UX [115]. Cultural differences present both opportunities and challenges in terms of how people interact with technology, consequently influencing UX [116]. Additionally, in a voice context, the dialect and accent of users have an impact on UX, and these aspects vary globally [117]. For example, scales developed to assess UX in the English-speaking countries may not effectively capture the experience of individuals, whose native language is not English and are not familiar with some English expression.

3) THE SCALES UX FACET MEASUREMENT INTENT

This study proceed to examine the UX aspects which each scale was developed to measure. Fig. 4, presents a heat map illustrating the distribution of the intended UX aspects measured by the scales. While it is evident that the scales are developed to capture either the holistic UX or its relatable aspect, it is important to identify the specific focus areas of each scale. The findings indicate that only 33% (n=7) of the scales are designed to assess the holistic concept of UX. The remaining scales target a specific aspect related to UX. Among the identified aspects, usability emerges as the most frequently measured aspect, with 38% (n=8) of the scales focusing on it. Some researchers have also employed the concept of "quality of use," which refers to how well the system fulfil its intended purpose and satisfies the user's needs within a specific context [118]. Voice quality is another prominent aspect of UX that is extensively measured, with 33% (n=7) of the scales dedicated to its assessment. Voice quality plays a vital role in voice interface systems like

VA [119]. In relation to other aspects of UX, the scales have been developed to measure enjoyability, affective response, anthropomorphism, user interaction, and personality, each accounting for approximately 5% (n=1) of the scales.

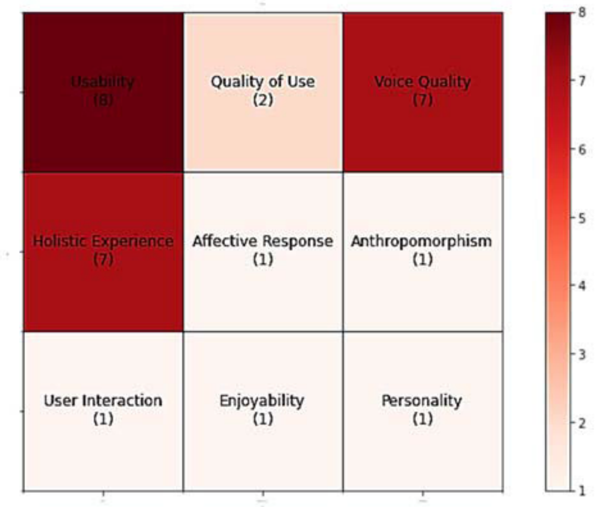

FIGURE 4. Heat map showing the different measurement aspects of the scales.

4) NOVELTY OF SCALES

The scales were grouped into two categories: novel and modified. A scale is categorized as novel if it presents and measures some new dimensions with a unique item list that is not wholly based on a particular previously developed scale [120]. A modified scale is created as an update of a previous existing scale (parent). Modified scales are typically developed through item reduction or addition [121]. The results suggest that some dimensions are more researched and popular than the others. Most researchers have focused on the hedonic aspect of the VA, which is evident from the maximum number of dimensions (n=11) being considered for the "pleasantness" tag. Compared to tags such as system speed that only has 1 dimension under it. These results align with existing HCI theories, which state that any information system is primarily used for utilitarian and hedonic purposes. This study identified the majority of the scales as novel, accounting for 57% (n=12), while the remaining 43% (n=9) were categorized as modified versions. Although the scale modifications were carried out for different reasons, the result shows that in most cases, there is at least an interval of a decade between an original scale and its modified version. For instance, [92] created SUISQ-R and SUISQ-MR, which are a modifications of SUISQ [90], which occurred after nearly a decade. Likewise, [96] created UEQ-S, which is a modified version of UEQ [95] roughly after a similar time frame.

5) STUDIES THAT UTILIZED DEVELOPED SCALE.

Numerous studies have employed scales to evaluate various aspects of voice assistants' user experience, contributing to a comprehensive understanding of user experience evaluation and the continuous improvement of conversational systems. Certain aspects of user experience can be applied across communication modality [104]. As earlier mentioned, some scales used by researchers for measuring the user experience of voice assistants, are not originally designed specifically for voice assistants nor voice user interface. For instance, UEQ and SUS were used by [25] to measure the usability, user experience, and usefulness of the Google Home smart speaker. Furthermore, Attrakdiff was used by [23] to investigate the relationship between task success and user experience of voice assistants, and [61] used it to compare the user experience of ambient voice assistants and button-activated voice assistants.

Some studies incorporate multiple scales to carry out their objectives. To illustrate, [37] compared the validity of SASSI, SUISQ-R, SUS, and Attrakdiff in measuring the user experience while using Alexa. Furthermore, [91] used SUS and SUISQ to measure the experience of a conversational agent to improve HPV vaccine coverage and reduce the burden of vaccine counseling for human providers. Some studies have taken the advantages of scale modification by using the modified scales instead. A Study by [89] opted to use the MOS-X version of MOS to measure the effect of the voice assistant's gender and communication style. Likewise, [94] employed SUISQ-MR, a modification of SUISQ, to measure the experience of heavy users and light users when using a VA that may affect their adoption. Some scales have not been adopted by other studies yet, which could be attributed to their recent development, or their low popularity.

Another important aspect is the understanding the diverse user groups and their specific needs when utilizing the scales. Studies have evaluated varied use cases - personal assistants [37], media centers [98], health counseling [91], automotive interfaces [122] etc. For instance [83] explores the usability and user experiences of speech-only interaction by elderly users using SASSI. The results indicate a high potential for speech-only interaction for elderly users in both indoor and outdoor environments, with an overall positive attitude and high acceptance. Another study by [94] showcased the importance of understanding the differences in difficulties faced by heavy and light users of voice user interfaces (VUIs). Also [107] highlighted the effect of prior experience on the user experience of voice assistant. Therefore, it is safe to say that understanding the diverse user groups and cases will satisfy the needs of all the user in adoption research, increment of inclusive accessibility, and provides practical insights for design improvements of conversational systems.

B. RQ$_2$: WHAT CONCEPTUAL DIMENSIONS ARE CURRENTLY PRESENT ON SCALES USED FOR MEASURING UX OF VA?

The concept of UX is multi-dimensional, and the specific dimensions vary depending on the measurement context.

TABLE 6. Dimension classification.

Tag Name	Dimensions	Scales
Intelligence The ability of responding to the user requests smartly and in an informed manner.	Response quality	UEQ+
	Intelligibility	UEQ+ MOS, MOS-X; MOS-R, MOS-R3
Accuracy The quality or state of being correct and/or precise.	Accurate/inaccurate	UEQ+
	System Response Accuracy	SASSI
	Effectiveness	UEXECS
	Error-free use	SUXES
Pleasantness The level of likeability, enjoyment and fun obtained after using the system.	Enjoyability	PUEVA
	Likeability	SASSI, CAS
	Unlikeable /Likeable/Unpleasant/Pleasant	UEQ+
	Boring /Entertaining	UEQ+
	Enjoyment /Fun	UEXECS, PUEVA
	Customer Service Behaviors	SUISQ, SUISQ-R, SUISQ-MR
	Entertainment	CAS
	Hedonic Quality	Attrakdiff, Attrakdiff2, UEQ-S
	Affect/Emotion	UEXECS
	Affective	VUS
	Pleasantness	SUXES
Efficiency Accomplishing the highest productivity by giving the least mental or physical effort.	Cognitive Demands	SASSI
	Efficiency	UEQ, UEXECS
	Attention	QOE
Irritability The level of annoyance and feelings of frustration with the system.	Annoyance	SASSI
	Verbosity	SUISQ, SUISQ-R, SUISQ-MR
Ease of use The easiness and effortlessness with which a system can be used and that which has minimum learning efforts.	User Goal Orientation	SUISQ, SUISQ-R, SUISQ-MR
	Complicated/ Simple/ Unambiguous/Ambiguous	UEQ+
	Recognition and Visibility	VUS
	Easiness	QOE
	Task Difficulty	CAS
	Clearness	SUXES
	Usable	SUS
	Learning curve	SUXES
	Learnable	SUS
System speed How fast the system responds to the user.	Speed	SASSI, SUXES
Synthetic Voice Naturalness The similarity of the system voice to the voice of a real-life human.	Speech Characteristics	SUISQ, SUISQ-R, SUISQ-MR
	Artificial/Natural	UEQ+
	Naturalness	MOS, MOS-X, MOS-R, MOS-R3, CAS, SUXES
	Fluency	MOS-R3
	Prosody	MOS-X
	Anthropomorphism	CAUS
Aptness The suitability and appropriateness of the system based on the user needs.	Habitability	SASSI
	Suitability	UEQ+
	Personalized Real-Time Interaction	CAUS
Usefulness Represents the value provided by the system in terms of being helpful and useful.	Useful	UEQ+, SUXES
	Helpful	UEQ+
	Helpfulness	CAS
Reliability The quality of being trustworthy or performing consistently well.	Robustness	SUXES
	Reliability	QOE
	Trust	CAS, CAUS
	Dependency	QOE
	Social Impression	MOS-x
	Dependability	UEQ
Aesthetics Relates to the looks and beauty of the product	Appeal	Attrakdiff2
	Attractiveness	UEQ
	Aesthetic/Appeal	UEXECS, Attrakdiff1
Engagement Relates to how the system encourages the user to keep using it	Presence	QOE
	Personality	PEUVA
	Stimulation	UEQ
	Involvement.	QOE
	Engagement /Flow	UEXECS
	Motivation	UEXECS
User Satisfaction	Satisfaction	QOE, UEXECS
	Future Use	SUXES

TABLE 6. ***(Continued.)*** **Dimension classification.**

It measures the amount of satisfaction that a user has experienced after using the products.		
Pragmatic Quality Relates to the functionality and practicality of the system.	Ergonomics	Attrakdiff1, Attrakdiff2
	Functionality	VUS, PUEVA, UEXECS
	Pragmatic Quality	UEQ-S

Thus, to address RQ$_2$ and understand the scales used for measuring UX of VA, it is important to identify and comprehend the dimensions the scales capture. Table 6 presents the classification of dimensions. It is important to note that some dimensions, although conceptually similar, may be named differently by different authors. In order to avoid redundancy during the data analysis, word classification of the dimensions was carried out based on the similarity in dimension definitions and were grouped into distinct tag names. Each tag name has been conceptually defined to encompass similar dimensions associated with it, distinguishing it from the other tags. From Table 6, 64 dimensions were highlighted in total across 21 scale. The 64 dimensions were grouped into 15 unique tags that are conceptually defined by the authors to avoid redundancy and represent the different UX dimension the current VA scales have covered. The results suggest that some dimensions are more researched and popular than the others. Most researchers have focused on the hedonic aspect of the VA, which is evident from the maximum number of dimensions (n=11) being considered for the "pleasantness" tag. Compared to tags such as system speed that only has 1 dimension under it. These results align with existing HCI theories, which state that any information system is primarily used for utilitarian and hedonic purposes. Moreover, unique tags are also identified, e.g., "intelligence" with (n=2) dimensions, "synthetic voice naturalness" with (n=6) dimensions, and "irritability" with (n=2) dimensions, among others. This indicates that the scope of UX for novel technologies like the conversational AI scenario is much broader.

C. RQ$_3$: WHAT METHODOLOGICAL AND OPERATIONALIZATION RIGOR WERE EMPLOYED WHILE DEVELOPING SCALES USED FOR MEASURING UX OF VA?

To answer RQ$_3$, the results were presented in Table 7.

These results are presented based on the evaluation and coding criteria that was previously mentioned in Table 3. Moreover, Table 7 also presents the overall percentage of the criteria reported when developing each of the scales. The study present 7 evaluation criteria. However, only 6 are considered for answering RQ$_3$ since the last criterion is used for explaining RQ$_4$. The maximum permissible score for each scale is 17, which is converted to a percentage score. Although the study give a percentage point to each scale, the study reaffirm that the objective behind RQ$_3$ is not to give score to papers based on their quality or to convey the superiority of some papers over others. Instead, the scores represent the level of in-depth thoroughness and the different kind of criteria they employed while developing their scales, which is an essential subject in a relatively new domain. None of the scales fulfill all six of the first criteria coding, as indicated in Table 7. The study by [10] has the highest score of 79% in terms of reporting carrying out the most criteria. On average, the scales reported to fulfill 46% of the criteria. Modified scales, such as those by [88] and [98], report to fulfill at least 20% of the criteria. Similarly, novel scales like the one by [32] reported to fulfill at least 26% of the evaluation criteria. The average reported criteria fulfillment score for scales developed after 2020 is 57%. Among the evaluation criteria, "conceptual dimension" is the most reported criteria to be carry out with 93% of the scales reporting it. Only three scales fail to report the conceptual definition of their identified dimensions. The results of the "item generation" criterion show that deductive methods are more reported to be carry out than inductive methods, with 95% (n=20) of the scales reporting deductive approaches, while 52% (n=11) reporting the inductive approach. Additionally, most scales that report an inductive approach also carry out the deductive methods. Regarding the "item refinement" criterion, expert evaluation is the most commonly reported technique at 48% (n=10), followed by pilot tests at 43% (n=9).

Pre-tests are the least reported method for item refinement, with only 38% (n=8) of the scales utilizing this approach. It is worth noting that only four scales (19%) report the use of all three methods of item refinement during their development. Regarding "sampling accuracy" only 19% (n=4) of the scales reported a sample size that satisfies the criteria outlined in Table 3. In comparison, 57% (n=12) of the scales reported sample size that partially fulfills this criterion. Regarding the sampling technique, none of the scales specifically reports the sampling technique used; however, some scales (roughly 43%, n=9) hints at the nature of the sample used, which conclude the probable sampling technique. For instance, [90], [98] states that all the participants were students from the same organization and available to the authors, which hints towards the use of convenience sampling. Likewise, [104] claims to have used an equal number of male and female participants to compare their average scores, which hints toward the use of stratified sampling.

Concerning "data factorability," only 19% (n=4) of the scale's reported Bartlett's test of Sphericity and the KMO values. Likewise, only 14% (n=3) of the scales uses a non-PCA-based approach for factor extraction, while an overwhelming 62% (n=13) of the scales used PCA-based approach. Only 1 study [10] uses multiple factor extraction

TABLE 7. Empirical analysis of the evaluation and coding criteria.

Evaluation Criteria	Coding	Attrakdiff 1	Attrakdiff2	CAS	CAUS	MOS	MOS-R	MOS-R3	MOS-X	PUEVA	QOE	SASSI	SUISQ	SUISQ-R	SUISQ-MR	SUXES	UEXECS	UEQ	UEQ+	UEQ-S	SUS	VUS
Conceptual Dimension	Dimension name	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Dimension definitions	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Item Generation	Deductive Method	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
	Inductive Method	1	0	0	1	1	0	0	1	1	1	1	1	0	0	0	0	1	0	0	1	1
Item Refinement	Pretest	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0	1
	Expert Evaluation	1	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	1	0	1	0	1
	Pilot testing	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0	0	1	0	1	1
Sampling Accuracy	Sample size	0	0.5	0	0.5	0.5	0.5	0.5	1	0	0	0.5	1	1	1	0.5	0.5	0.5	0.5	0.5	0	0.5
	Sampling technique	0	0	0.5	0	0.5	0	0.5	0.5	0	0.5	0.5	0.5	0	0	0	0.5	0.5	0	0	0	0
Data Factorability	Correlation Matrix	0	1	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	1
	Bartlett's test of sphericity	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	Kaiser-Meyer-Olkin (KMO)	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Factor Analysis (EFA)	Factor Extraction	0.5	0.5	1	1	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0	0	0.5	0.5	0.5	0	1
	Factor Retention	0	0	0	1	1	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1	1
	Factor Rotation	1	1	1	0	1	0	0	0	1	0	1	1	1	1	0	0	1	1	1	1	1
Factor Analysis (CFA)	Dedicated samples	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Model Fit Indices	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
% of criteria reported		38	35	56	74	53	24	20	35	50	56	68	65	38	38	26	35	50	47	35	47	79

method as a part during EFA (PCA, unweighted least squares, and maximum likelihood). About 52% (n=11) of the scales reports their factor retention methods, with around (n=8) of them using the Scree plot technique and the remaining of the scales (n=3) using parallel analysis. Regarding CFA, none of the scales reported using a dedicated sample, while only 5% (n=1) of the scales report at least 1 model-fit indices.

D. RQ$_4$: WHAT ARE THE REPORTED RELIABILITY AND VALIDITY OF THE CURRENT SCALES USED FOR MEASURING UX OF VA?

For answering RQ4, the study analyze the seventh evaluation criterion as outlined in Table 3. The score points presented in Table 8 do not establish one validity as superior to another. Instead, they indicate whether the scale has reported the validity and reliability assessment. The maximum score a scale can achieve for fully reporting a construct validity is 2, which is obtained when the scale reports both types of construct validity. Therefore, if only one type of construct validity is reported or if the type of construct validity is not specified, a score of 1 is assigned. The same scoring method applies to criterion validity. Therefore, if a scale reports all types of validity (both types of constructs and criterion validity), content validity, and reliability, as per the criteria in Table 3, the full score would be 6 point, which is converted to 100 percent. About 38 % of the scales (n=8) reported the content validity measurements.

For construct validity, 48% of the scales (n=10) reports either convergent or discriminant validity. Only 1 scale reported both types of construct validity [103]. Regarding criterion validity, none of the scales report concurrent validity form of criterion validity, while only 24% of the studies (n=5) reported the criterion validity without specifying. Regarding the internal consistency of the scales, 57% of the studies (n=12) all reported dimensions which have Cronbach's alpha values above 0.7.

However, 24% of the studies (n=5) report some dimensions with less than 0.7 Cronbach's alpha. Finally, 19 % of the scales (n=4) did not report their reliability Cronbach's alpha at all. 81% of the scales (n=17) reported reliability and 71% of the scales (n=15) reported at least one type of validity. On an average a scale report about 30% of the reliability and

TABLE 8. Analyzing the reliability and validity of the scales.

Evaluation Criteria	Coding	Attrakdiff1	Attrakdiff2	CAS	CAUS	MOS	MOS-R	MOS-R3	MOS-X	PUEVA	QOE	SASSI	SUISQ	SUISQ-R	SUISQ-MR	SUXES	U2XECS	UEQ	UEQ+	UEQ-S	SUS	VUS
Validity	Content Validity	0	0	0	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1
	Construct Validity	0	1	2	0	1	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	0
	Criterion Validity	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0
Reliability	Internal Consistency	1	1	1	1	1	1	1	1	0	0	0.5	1	0.5	0.5	0	0	0.5	1	0.5	1	1
Score of Each Scale (Converted to 100)		17	33	50	33	67	33	33	33	17	0	25	67	58	58	0	0	25	17	9	50	33

validity measures during development. Moreover, none of the scales reported all the types of validity and reliability during development.

VI. DISCUSSION

In this study, the essence of UX in the context of VA and similar voice user interfaces is portrayed. While UX is not a new concept, it still remains crucial to investigate it, particularly in light of emerging human-to-machine (H2M) interaction, such as those found in VA communication. Therefore, this study seek to gain a comprehensive understanding of the different aspects of UX by focusing on the existing scales available in the research community used to measure the UX of the VA, together also delve into the detailed development of these scales to further comprehension.

A. UX SCENARIO IN VA PARADIGM

The detailed synthesis of the current scales yields 15 tags related to the UX of VA (Table 6). While some of these have been used for decades in UX research by the HCI community, e.g., ease of use, efficiency, or user satisfaction, other tags like aptness or engagement are relatively new. In order to understand whether the UX concepts outlined in Table 6 can be related to some already existing UX frameworks, the study adopted the framework by [18] as the reference for discussion. The framework is based on the V (Value), U (Usability), D (Desirability), and A (Adoptability) principles. The framework is chosen due to its simplicity, comprehensiveness, and previously used in multiple studies [123], [124], [125]. The study aims to categorize the dimension tags identified in Fig. 5 based on the VUDA principle. Value represents a crucial facet of UX, closely tied to a system's functionality and features. Accordingly, the value aspect encompasses pragmatic quality, reliability, and aptness of the VA, as these tags collectively assess the VA's functionality, suitability, and consistent performance. The second facet is usability, a prominent element in UX research. Usability pertains to the ease of using technology, encompassing the cognitive load necessary to complete tasks accurately with the VA. Consequently, the tags of ease of use, efficiency, accuracy, system speed, and satisfaction also address usability, as they are all connected to the VA's ease, accuracy, and efficiency in creating a satisfying UX. The desirability aspect is associated with users' emotional appeal and engagement, as well as the enjoyment and usefulness derived from using a system. Four related tags, namely pleasantness, usefulness, aesthetics, and engagement, have been grouped under desirability.

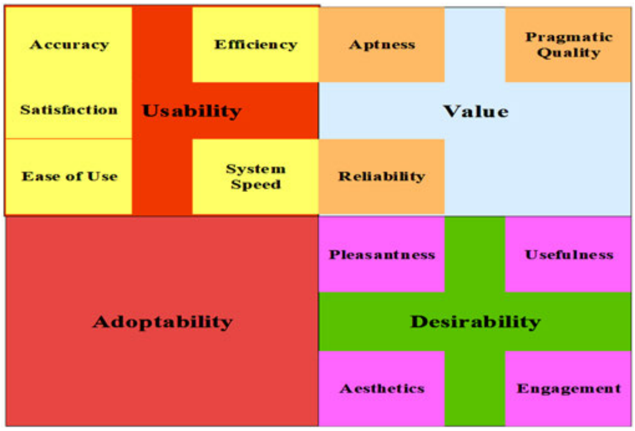

FIGURE 5. Classification of conceptually dimension tags using GUO UX framework.

Value, usability, and desirability collectively represent UX aspects that manifest after using a system. Conversely, the final aspect, adoptability, pertains to the UX before product use. It ensures that a specific product or system has been designed to align with users' natural curiosity of exploring its features. Unfortunately, none of the identified tags fit into this category. Following the classification of tags, three remained unclassified: intelligence, irritability, and synthetic voice naturalness, primarily because their conceptual meanings did not align with the selected framework. For instance, both intelligence and irritability are linked to the VA's personality.

Since VAs are AI-powered devices, they can respond intelligently to user requests, potentially triggering feelings of annoyance and frustration. Similarly, the tag of synthetic voice naturalness relates to the VA's voice features, their conversational style, and their similarity to a human voice. This metric also measures the strength of the bond that develops between humans and VAs, touching upon the concept of anthropomorphism, which is a central aspect of voice-based systems. Yet, traditional frameworks like [18] cannot capture all the dimension since it was created for GUI tools.

B. PROPOSED MODIFICATION TO THE UX FRAMEWORK

The limitations of the existing UX framework in classifying VA dimensions become apparent in the preceding discussion. Notably, there was no categorization of dimensional tags under adoptability, despite its fundamental role in the realm of UX. This omission may explain why it remains unclear why several users discontinue using VAs after an initial period of usage. Consequently, this study introduces an enhanced framework, as depicted in Fig. 6, which is better suited for comprehensively assessing the UX of VAs and other voice user interface systems. To measure the adoptability aspect, the study proposes three measures: a straightforward onboarding process, knowledge empowerment, and personalized service. When users initiate their interaction with a VA, the onboarding process should be uncomplicated, devoid of confusion or misunderstandings. One way this can be achieved is by using a "wake-up" word for the VA that is easy to pronounce and remember. Complex "wake-up" words that are difficult to pronounce may create confusion and lead to abandonment. The VA must be highly interactive in this initial stage and ensure to ask for follow-up questions to clarify doubts. Likewise, during the initial usage period, the VA may reward the users by using positive, conversational phrases when they complete some essential steps in the on boarding process to give them confidence and provide a pleasant experience.

Being AI-powered devices, the VA have several capabilities ranging from performing simple to complex tasks. Empowering the users with proper knowledge is, therefore, critical to let them know what features are available and what the VA can do and what to expect. The VA may provide quick tips in the form of actionable messages wherever appropriate or even enquire about the general well-being of the users if the users do not interact for a specific time period. Thus, the adoptability of the VA can be ensured by continuously updating the voice experience received by the users based on their usage patterns and creating an informed feedback loop.

Using AI and machine learning on the voice data that users generate, the VA can provide personalized recommendations and customizations to provide better UX. The current scales need to include this adoptability aspect that helps to create positive initial experiences. Consequently, another drawback of the Guo UX framework towards the VA conceptually tag dimensions classification is that it does not consider novel concepts like the synthetic voice naturalness.

Careful consideration of the related items of this tag reveals that the accent, style of speech, conversational mannerisms, use of filler words in between conversations, pitch, prosody, and rhythm, i.e., the various socio-linguistics characteristics of the voice of the VA make up this tag. The purpose behind measuring all these socio-linguistic characteristics is to check how far the VA can mimic humanness. In other words, to measure anthropomorphism which can alleviate the UX by invoking perceptions of likeliness. Friendliness and camaraderie. For UX frameworks that have been formulated when technology was perceived to be non-human/lifeless, it is evident that they fail to capture the human-likeliness aspect that is an integral part of UX measurement of anthropomorphic technologies like VA. Hence, the study propose that this additional aspect of anthropomorphism be included as a core UX measure.

The final drawback of the Guo UX framework in measuring UX of VA is it does not consider the intelligence and irritability aspects that current VA scales consider. Although intelligence and irritability are human-like attributes that can be classified under the anthropomorphism category, as explained above, yet these are conceptually different from anthropomorphism. For instance, Google search can remember a user's previous search history, and auto-fill various fields during subsequent searches since it learns from the previous search behaviors. Thus, Google search exhibits intelligence, but it is non-anthropomorphic unlike models such as ChatGPT which exhibit anthropomorphic characteristic. Moreover, a robotic toy can ideally look like a human and even imitate specific human actions like smiling, walking, or sitting but does not have any interaction capability, meaning that it is highly anthropomorphic but has limited intelligence. Therefore, the study propose anthropomorphism and intelligence as separate concepts that depend on the technology may be present at varying levels. More importantly, intelligence and irritability can be related to human personality traits suggesting that machine (VA) personality is essential to consider. Designing VA having the right personality for the users will help to improve the overall UX. Consequently, personality aspect should be included in the proposed UX framework.

C. COMMONALITY BETWEEN VA UX DIMENSIONS AND VA USER ACCEPTANCE (UA) DIMENSIONS

As mentioned previously, there is a strong connection between UX and User Acceptance, especially in the context of Voice Assistants (VAs), where acceptance plays a pivotal role in their adoption. This part investigate the relationship between the identified UX dimensions and the established User Acceptance models in the field of voice assistant research. *First*, when users perceive a system as intelligent, precise, and effective in fulfilling their needs (related to intelligence and accuracy in UX dimensions), this closely aligns with dimensions found in VA acceptance models, such as Information Quality, intelligence, accuracy, and Perceived Usefulness [126]. Systems that exhibit these qualities are generally more likely to be accepted. *Second*, when a system is efficient and user-friendly (related to efficiency and ease of use in UX dimensions), it corresponds to elements in voice assistant acceptance models, such as Perceived Ease of Use and effort expectancy, which contribute to higher acceptance [127]. *Third*, when the UX is enjoyable, engaging,

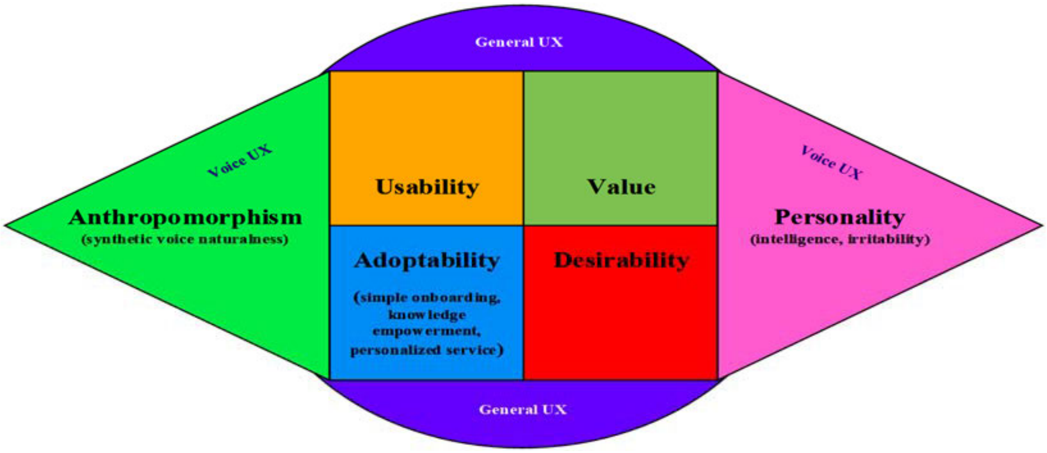
General UX
Voice UX
Anthropomorphism
(synthetic voice naturalness)
Usability
Value
Adoptability
(simple onboarding, knowledge empowerment, personalized service)
Desirability
Voice UX
Personality
(intelligence, irritability)
General UX

FIGURE 6. Proposed modified UX framework for the voice assistants.

and pleasant, it has a direct positive impact on voice assistant acceptance dimensions, such as attitude [126] and perceived enjoyment [16], ultimately fostering greater user acceptance. Conversely, negative aspects of the UX, such as irritability, can adversely affect user satisfaction, which is crucial for acceptance. Furthermore, the naturalness of synthetic voices influences perceived quality, which in turn shapes overall user acceptance. The reliability of a system indicates trustworthiness and consistent performance, aligning with the trust factor in voice acceptance models, thus enhancing overall acceptance [127]. *Lastly*, aesthetics and engagement significantly contribute to product features, which are closely linked to voice assistant user acceptance [127]. Aptness closely relates to the completeness factor within the voice assistant user acceptance model [126].

In conclusion, these identified UX dimensions are closely intertwined with the dimension of user acceptance. They shape users' perceptions, attitudes, and satisfaction, all of which constitute crucial elements in numerous User Acceptance models. This underscores the concept that a favorable UX, characterized by these dimensions, typically results in greater user acceptance and the continued adoption of voice assistant technology.

D. UNRAVELING THE SCALE DEVELOPMENT EMPLOYED CRITERIA

The overall results show that studies did not consistently report carrying out all the evaluation criteria, reaffirming the notion that the criteria employed are subjective and dependent on the developer. This means that, even though the requirements are essential, they are not strictly adhered to. The discussion is presented on a case-by-case basis. Each of the studies employs the first criterion of the conceptual dimension, which is expected as it signifies the primary purpose of the scale. However, sometimes authors fail to provide adequate definitions of the dimension. One common reason is that some dimension names are dictionary-based, making further explanations unnecessary. Another reason is that modified scales often use the same definition as their parent scales, rendering a new description superfluous. However, this flaw might become problematic if the scales were to be split into further sub-dimensions in the future. Despite every scale addressing the conceptual dimension, many terms used are synonyms, resulting in very similar items. For instance, in [32], the happiness dimension is similar to the entertainment dimension in the CAS scale [103]. This redundancy could have been avoided through expert review and intensive literature research during the conceptual dimension stage, which would improve the content validity of the scales. Many studies employ literature reviews as part of deductive methods for creating scale items. However, the choice of deductive methods limits the scope of creativity when constructing the items. Moreover, since deductive methods rely on existing literature and previous scales, they should primarily be used when authors intend to create modified versions of the scales. Instead, inductive methods should be employed more frequently in case of new novel scales.

Content validity is significantly affected by the number of items on a scale; therefore, modified scales often opt for item reduction by eliminating dimensions to increase their content validity. In cases of item reduction, a deductive method should be employed for principal component selection [92].

Many studies tend to overlook the reporting of the Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity, neglecting to emphasize the data's suitability for factor analysis. However, recognizing inherent correlations among variables can save authors valuable time by indicating the appropriateness of the data for Exploratory Factor Analysis (EFA). Factor analysis constitutes a fundamental aspect of scale development. While some studies exclusively perform EFA and others combine EFA and Confirmatory Factor Analysis (CFA), however, conducting both concurrently is considered a best practice. Novel studies should incorporate EFA into their factor analysis methodologies, as EFA uncovers fresh correlations within factor structures. Moreover,

TABLE 9. Scale comparison in their criteria execution.

Criteria	Novel Studies	Modified studies
Conceptual Dimension	Conceptual dimension and dimension definition are rigorously carried out.	Conceptual dimensions are inherited from parent scale. New dimension definitions are weak if carried out at all
Item Generation	Both deductive and inductive methods are employed.	Largely deductive methods as literature review of all versions of parent scale
Pretest	Decently carried out	Process is weakly considered if at all.
Pilot test	Decently carried out	Mildly carried out on scale that are intending to create a new additional dimension from the parent scales
Expert Evaluation	Decently carried out, the process relates to content validity	Decently carried out, the process relates to content validity
Sampling Accuracy	The sample size rule of 10:1, and > 300 was very weakly employed. However, majority of the studies sample size were greater than 30	The sample size rule of 10:1, and > 300 was fairly employed. However, majority of the studies sample size were greater than 30
Correlation Matrix	Mildly carried out	Weakly carried out
Kaiser-Meyer-Olkin (KMO)	Weakly carried out	Not carried out
Principal Component Analysis (PCA)	Mildly carried out	Carried out by all the studies
Maximum Likelihood	Weakly carried out	Not carried out
Principal Axis Factoring	Weakly carried out	Not carried out

when scales aspire to achieve high reliability and validity, both EFA and CFA should be conducted. Scales that have been translated from another language, such as the UEQ, are considered new due to alterations in their word structure during translation, impacting item correlations and necessitating item removal. Principal Component Analysis (PCA) might not always be the most suitable method for latent construct identification; however, it remains widely utilized for item reduction, as it identifies the primary component items within the parent scale. Another crucial consideration is that CFA and EFA should be performed on separate samples, necessitating the division of data into two subsamples: one for EFA and the other for CFA. When both EFA and CFA are executed on the same sample, it verifies the data rather than the scale itself.

Content validity plays a pivotal role in forming the initial impression of a scale due to its ease and promptness. Experts, possessing a deeper understanding of the measures are often considered the best method for increasing content validity. Construct validity is the most frequently reported validity in studies, owing to its objective verifiability, making it a crucial component of psychometric properties. In certain cases, construct validity is applied to modified scales, such as [92], to ensure no overfitting in the reused sample. Internal consistency, typically measured by alpha values ranging between 0.8 and 0.9, is frequently employed when developing a scale. Additionally, test-retest reliability should be considered when adapting previous scales to accommodate the latest advanced VA systems.

E. VARIANCES OF CRITERIA EMPLOYED IN NOVEL AND MODIFIED SCALES

As mentioned earlier, each criterion in scale development holds significance. However, the selection and utilization of criteria depend on the author's rationale and purpose for developing the scale.

To provide future researchers with a concise overview of the criteria reported during the development of both novel and modified scales used for measuring UX of VA, Table 9 is presented. This synopsis aims to offer researchers insights into what to consider when developing either a new or modified scale. It is important to note that the information is specifically based on the analysis of the 21 scales used in this study. The terms "decently," "mildly," and "weakly" were used and employed during the overview. "Decently" indicates that the criterion was reported a significant amount of the time, but not always. "Mildly" suggests that the criterion was reported on average. "Weakly" implies that the criterion was reported minimally or not reported at all.

F. STRATEGIES FOR ENSURING THE SCALES' ADAPTABILITY TO TECHNOLOGICAL ADVANCEMENTS

In the rapidly evolving landscape of Voice Assistant (VA) technology, the continuous advancement necessitates ongoing updates to scales to ensure their relevance. Therefore, collaborations between stakeholders which include researchers, industry stakeholders and technology developers will deem beneficial in keeping up with the technological advancement of VA. Moreover, Implementation of a structured feedback mechanism where users, developers, and researchers can provide insights on the performance and limitations of existing scales in current technology will also prove beneficial. Furthermore, encouraging open communication to capture real-world experiences and challenges encountered while using current scales is also vital. Another way for scales to keep up with technological advancement is to prioritize a user-centered design approach, involving more end-users in the evaluation and refinement of scales to ensure their practical utility and relevance.

VII. CONCLUSION AND FUTURE WORKS

As the ubiquity of voice assistants (VA) continues to increase, more people are engaging with them daily. Voice interfaces, in general, will play a significant role in future interactive systems. A well-designed UX measurement will contribute significantly to the overall improvement of voice-based interfaces. In this literature review, we identified 21 scales related to voice UX. The majority of these scales were developed in a Western context. Traditional UX frameworks need to evolve to capture and explain the different dimensions of UX in the VA scenario, given the fundamental differences between GUI systems and VAs. However, there is a stark contrast between the development of novel and modified scales. For instance, modified scales overwhelmingly employ the PCA method during the factor analysis phase, whereas novel scales sometimes opt for other methods, such as maximum likelihood and principal axis factoring. Concerning validity, construct validity was the most frequently reported type. However, internal consistency has been reported by almost all the studies, except for a few. This suggests that studies often report their study's reliability but leave the validity somewhat vague. Therefore, it's important to report validity measures for the scales more comprehensively.

A. LIMITATION

This study is not without its limitations. Firstly, all the scales considered in the study are developed in English. Therefore, other scales developed in different languages from non-English speaking countries are not included. This can be a setback because majority of non-English speaking countries are technologically advanced and therefore might already have their own scales used for measuring UX of VA. *Secondly*, the points attributed to the studies during their development criteria are not reflective of the thoroughness of the study but are a direct result of what the author's decide to report in their scale development study. Therefore, a low point demonstrates the author's failure to report the coding criteria or an aspect of the psychometric properties (validity and reliability) they carried out. *Thirdly*, the field of voice assistants' user experience has garnered significant attention, leading to the undertaking of studies on scales. This study recognizes the absent of some newer scales and the exclusion of studies that have standardized scales more closely associated with chatbot. This limitation might exclude some relevant scales.

B. FUTURE WORK RECOMMENDATION

Based on the review study conducted, the researchers recommend the identification of future work recommendations on the aspect of UX of VA. These will have the potential to expand the horizons of knowledge and deliver substantial value to the HCI community. These recommendations include: *First*, anthropomorphism and the concept of machine personality represent emerging and crucial facets within the voice paradigm. To gain a deeper understanding of how these elements impact the UX of voice assistants, including potential considerations related to the uncanny valley, future research should place significant emphasis on conducting extensive investigations. *Second*, while certain surface-level similarities and relationships between UX and user acceptance models in the context of voice assistants are identified, there is still much more to discover when delving into the realm of User Acceptance for voice assistants. Future work should entail a comprehensive exploration of the dimensions of User Acceptance, including their development and how they interplay with UX. *Third*, as demonstrated by [42], the integration of traditions and culture into both the system model (which includes technical aspects like hardware) and the user model (encompassing mental models, expectations, and cultural backgrounds) can significantly impact UX. Moreover, there is a substantial disparity in the origins of current scales used for scales used for measuring UX of VA, with Africa and Australia being the least originators. Given this diversity in scale origins, it is imperative to understand how culture can influence UX in VA context, especially considering that each location possesses its unique cultural characteristics. For instance, in voice context, the dialect and accent of users have an impact on UX, and these aspects vary globally; this raises the question of whether a scale developed in one location can accurately capture the true UX in another location. Consequently, there is a need for a prospective study to evaluate how the incorporation of cultural elements into each of these models can enhance the UX in the context of voice assistants. *Fourth,* Given the popularity and innovation of ChatGPT, chatbot are surpassing the capabilities of traditional conversational agents [128]. This shift has heightened attention on the text-to-speech (TTS) methodology. Subsequent research should utilize standardize chatbot scales such as (BUS) [129] which is available in multiple language for measuring the user experience of text to speech chatbot elements. This approach will contribute to emphasizing the relevance of using standardized chatbot scales in measuring voice form of conversational agents.

REFERENCES

[1] IMDb. (1985). *Back to the Future (1985)—-IMDb*. Accessed: Oct. 2, 2022. [Online]. Available: https://www.imdb.com/title/tt0088763/

[2] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan, "What makes a good conversation: Challenges in designing truly conversational agents," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–12, doi: 10.1145/3290605.3300705.

[3] R. Guerrieri and M. Kokinova, "Does instruction in the use of personal digital assistants increase medical students' comfort and skill level?" *Med. Reference Services Quart.*, vol. 28, no. 1, pp. 33–43, Jan. 2009, doi: 10.1080/02763860802615955.

[4] G. H. Walker, N. A. Stanton, and M. S. Young, "Where is computing driving cars?" *Int. J. Hum.-Comput. Interact.*, vol. 13, no. 2, pp. 203–229, Jun. 2001, doi: 10.1207/s15327590ijhc1302_7.

[5] J. Brinkley, B. Posadas, I. Sherman, S. B. Daily, and J. E. Gilbert, "An open road evaluation of a self-driving vehicle human–machine interface designed for visually impaired users," *Int. J. Hum.-Comput. Interact.*, vol. 35, no. 11, pp. 1018–1032, Jul. 2019, doi: 10.1080/10447318.2018.1561787.

[6] J. Kim, K. Merrill, K. Xu, and D. D. Sellnow, "My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 20, pp. 1902–1911, Dec. 2020, doi: 10.1080/10447318.2020.1801227.

[7] L. Qiu and I. Benbasat, ''Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars,'' *Int. J. Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 75–94, Sep. 2005, doi: 10.1207/s15327590ijhc1901_6.

[8] Y. Zhu, M. Janssen, R. Wang, and Y. Liu, ''It is me, chatbot: Working to address the COVID-19 outbreak-related mental health issues in China. User experience, satisfaction, and influencing factors,'' *Int. J. Hum.-Comput. Interact.*, vol. 38, no. 12, pp. 1182–1194, Jul. 2022, doi: 10.1080/10447318.2021.1988236.

[9] Federica Laricchia. (2019). *Number of Voice Assistants in Use Worldwide 2019-2024 | Statista*. Accessed: Oct. 25, 2022. [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[10] D. S. Zwakman, D. Pal, and C. Arpnikanondt, ''Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa,'' *Social Netw. Comput. Sci.*, vol. 2, no. 1, p. 28, Feb. 2021, doi: 10.1007/s42979-020-00424-4.

[11] M. Hassenzahl and N. Tractinsky, ''User experience—A research agenda,'' *Behaviour Inf. Technol.*, vol. 25, no. 2, pp. 91–97, Mar. 2006, doi: 10.1080/01449290500330331.

[12] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. S. Vermeeren, and J. Kort, ''Understanding, scoping and defining user experience,'' in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2009, pp. 719–728, doi: 10.1145/1518701.1518813.

[13] A. G. Mirnig, A. Meschtscherjakov, D. Wurhofer, T. Meneweger, and M. Tscheligi, ''A formal analysis of the ISO 9241–210 definition of user experience,'' in *Proc. 33rd Annu. ACM Conf. Extended Abstr. Human Factors Comput. Syst.*, Apr. 2015, pp. 437–446, doi: 10.1145/2702613.2732511.

[14] J. A. Bargas-Avila and K. Hornbæk, ''Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience,'' in *Proc. Conf. Hum. Factors Comput. Syst.*, 2011, pp. 2689–2698, doi: 10.1145/1978942.1979336.

[15] D. Mugo, K. Njagi, B. Chemwei, and J. Motanya, ''The technology acceptance model (TAM) and its application to the utilization of mobile learning technologies,'' *Brit. J. Math. Comput. Sci.*, vol. 20, no. 4, pp. 1–8, Jan. 2017, doi: 10.9734/bjmcs/2017/29015.

[16] C.-M. Chao, ''Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model,'' *Frontiers Psychol.*, vol. 10, pp. 1–14, Jul. 2019, doi: 10.3389/fpsyg.2019.01652.

[17] K. Hornbæk and M. Hertzum, ''Technology acceptance and user experience: A review of the experiential component in HCI,'' *ACM Trans. Comput.-Hum. Interact.*, vol. 24, no. 5, pp. 1–30, Oct. 2017, doi: 10.1145/3127358.

[18] G. Frank. (2022). *More Than Usability: The Four Elements of User Experience, Part I: UXmatters*. Accessed: Aug. 8, 2022. [Online]. Available: https://www.uxmatters.com/mt/archives/2012/04/more-than-usability-the-four-elements-of-user-experience-part-i.php

[19] A. B. Kocabalil, L. Laranjo, and E. Coiera, ''Measuring user experience in conversational interfaces: A comparison of six questionnaires,'' in *Proc. Electron. Workshops Comput.*, Jul. 2018, pp. 1–12, doi: 10.14236/ewic/hci2018.21.

[20] I. Díaz-Oreiro, G. López, L. Quesada, and L. Guerrero, ''Standardized questionnaires for user experience evaluation: A systematic literature review,'' in *Proc. 13th Int. Conf. Ubiquitous Comput. Ambient Intell.*, Nov. 2019, p. 14, doi: 10.3390/proceedings2019031014.

[21] F. Habler, M. Peisker, and N. Henze, ''Differences between smart speakers and graphical user interfaces for music search considering gender effects,'' in *Proc. 18th Int. Conf. Mobile Ubiquitous Multimedia*, Nov. 2019, pp. 1–7, doi: 10.1145/3365610.3365627.

[22] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu, ''Patterns for how users overcome obstacles in voice user interfaces,'' in *Proc. Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–7, doi: 10.1145/3173574.3173580.

[23] M. Kurz, B. Brüggemeier, and M. Breiter, ''Success is not final; Failure is not fatal-task success and user experience in interactions with Alexa, Google Assistant and Siri,'' in *Proc. Int. Conf. Human-Comput. Interact.* (Lecture Notes in Computer Science), vol. 12764, 2021, pp. 351–369, doi: 10.1007/978-3-030-78468-3_24.

[24] D. Ghosh, P. S. Foong, S. Zhang, and S. Zhao, ''Assessing the utility of the system usability scale for evaluating voice-based user interfaces,'' in *Proc. ACM Int. Conf.*, Apr. 2018, pp. 11–15, doi: 10.1145/3202667.3204844.

[25] A. Pyae and T. N. Joelsson, ''Investigating the usability and user experiences of voice user interface: A case of Google home smart speaker,'' in *Proc. 20th Int. Conf. Human-Comput. Interact. Mobile Devices Services Adjunct*, Sep. 2018, pp. 127–131, doi: 10.1145/3236112.3236130.

[26] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, ''Best practices for developing and validating scales for health, social, and behavioral research: A primer,'' *Frontiers Public Health*, vol. 6, p. 149, Jun. 2018, doi: 10.3389/fpubh.2018.00149.

[27] F. F. R. Morgado, J. F. F. Meireles, C. M. Neves, A. C. S. Amaral, and M. E. C. Ferreira, ''Scale development: Ten main limitations and recommendations to improve future research practices,'' *Psicologia, Reflexão e Crítica*, vol. 30, no. 1, pp. 1–20, Jan. 2018, doi: 10.1186/s41155-016-0057-1.

[28] S. Borsci, A. Malizia, M. Schmettow, F. van der Velde, G. Tariverdiyeva, D. Balaji, and A. Chamberlain, ''The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents,'' *Pers. Ubiquitous Comput.*, vol. 26, no. 1, pp. 95–119, Feb. 2022, doi: 10.1007/s00779-021-01582-9.

[29] M. K. Othman, A. Nogoibaeva, L. S. Leong, and M. H. Barawi, ''Usability evaluation of a virtual reality smartphone app for a living museum,'' *Univers. Access Inf. Soc.*, vol. 1, pp. 1–18, May 2021, doi: 10.1007/S10209-021-00820-4.

[30] S. A. Jin, ''The impact of 3D virtual haptics in marketing,'' *Psychol. Marketing*, vol. 28, no. 3, pp. 240–255, Mar. 2011, doi: 10.1002/mar.20390.

[31] P. Moreno-Ger, J. Torrente, Y. G. Hsieh, and W. T. Lester, ''Usability testing for serious games: Making informed design decisions with user data,'' *Adv. Hum.-Comput. Interact.*, vol. 2012, pp. 1–13, Jan. 2012, doi: 10.1155/2012/369637.

[32] S. Li, S. Krome, I. Mandel, M. Walch, and W. Ju, ''The PUEVA inventory: A toolkit to evaluate the personality, usability and enjoyability of voice agents,'' 2021, *arXiv:2112.10811*.

[33] M. J. Hilsenroth, M. D. Blagys, S. J. Ackerman, D. R. Bonge, and M. A. Blais, ''Measuring psychodynamic-interpersonal and cognitive-behavioral techniques: Development of the comparative psychotherapy process Scale,'' *Psychotherapy, Theory, Res., Pract., Training*, vol. 42, no. 3, pp. 340–356, Sep. 2005, doi: 10.1037/0033-3204.42.3.340.

[34] J.-C. Pillet, C. Vitari, F. Pigni, and K. Carillo. (2018). *Detecting Biased Items When Developing a Scale: A Quantitative Method*. Accessed: Apr. 24, 2022. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-01923612

[35] J. Sauro. (2016). *The Challenges and Opportunities of Measuring the User Experience*. Accessed: Nov. 1, 2022. [Online]. Available: https://uxpajournal.org/challenges-opportunities-measuring-user-experience/

[36] A. Berdasco, G. López, I. Diaz, L. Quesada, and L. A. Guerrero, ''User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana,'' *UCAml*, vol. 2019, p. 51, Nov. 2019, doi: 10.3390/PROCEEDINGS2019031051.

[37] B. Brüggemeier, M. Breiter, M. Kurz, and J. Schiwy, ''User experience of Alexa when controlling music: Comparison of face and construct validity of four questionnaires,'' in *Proc. 2nd Conf. Conversational User Interfaces*, Jul. 2020, pp. 1–9, doi: 10.1145/3405755.3406122.

[38] P. Phichai, J. Williamson, and M. Barr, ''Alternative design for an interactive exhibit learning in museums: How does user experience differ across different technologies-VR, tangible, and gesture,'' in *Proc. 7th Int. Conf. Immersive Learn. Res. Network*, Aug. 2021, pp. 1–8, doi: 10.23919/iLRN52045.2021.9459414.

[39] A. Følstad and C. Taylor, ''Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues,'' *Qual. User Exper.*, vol. 6, no. 1, pp. 1–17, Dec. 2021, doi: 10.1007/s41233-021-00046-5.

[40] N. Kaewpanukrangsi and C. Kerdvibulvech, ''Gesture-based communication via user experience design: Integrating experience into daily life for an arm-swing input device,'' *Augmented Hum. Res.*, vol. 4, no. 1, p. 2, Dec. 2019, doi: 10.1007/s41133-019-0013-6.

[41] L. Graichen. (2020). *Gestures for Human-machine Interaction. Design Aspects, User Experience and Impact on Driving Safety*. [Online]. Available: https://nbn-resolving.org/urn

[42] R. Long, X. Liu, T. Lei, X. Chen, and Z. Jin, ''The impact of Chinese traditional cultural on the gesture and user experience in mobile interaction design,'' in *Proc. Int. Conf. Cross-Cultural Design* (Lecture Notes in Computer Science), vol. 10281, 2017, pp. 49–58, doi: 10.1007/978-3-319-57931-3_5.

[43] R. Rohan, D. Pal, and S. Funilkul, ''Hey Alexa…examining factors influencing the educational use of AI-enabled voice assistants during the COVID-19 pandemic,'' in *Proc. 15th Int. Conf. Knowl. Smart Technol. (KST)*, Feb. 2023, pp. 1–6, doi: 10.1109/KST57286.2023.10086856.

[44] M. Milne-Ives, C. de Cock, E. Lim, M. H. Shehadeh, N. de Pennington, G. Mole, E. Normando, and E. Meinert, ''The effectiveness of artificial intelligence conversational agents in health care: Systematic review,'' *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e20346, doi: 10.2196/20346.

[45] K. Denecke and R. May, ''Usability assessment of conversational agents in healthcare: A literature review,'' in *Challenges of Trustable AI and Added-Value on Health*. U.K.: IOS Press, May 2022, pp. 169–173, doi: 10.3233/SHTI220431.

[46] R. Boumans, Y. van de Sande, S. Thill, and T. Bosse, ''Voice-enabled intelligent virtual agents for people with amnesia: Systematic review,'' *JMIR Aging*, vol. 5, no. 2, Apr. 2022, Art. no. e32473, doi: 10.2196/32473.

[47] F. L. I. Dutsinma, D. Pal, S. Funilkul, and J. H. Chan, ''A systematic review of voice assistant usability: An ISO 9241–11 approach,'' *Social Netw. Comput. Sci.*, vol. 3, no. 4, p. 267, Jul. 2022, doi: 10.1007/s42979-022-01172-3.

[48] C. Tubin, J. P. M. Rodriguez, and A. C. B. de Marchi, ''User experience with conversational agent: A systematic review of assessment methods,'' *Behav. Inf. Technol.*, vol. 41, no. 16, pp. 3519–3529, 2021, doi: 10.1080/0144929X.2021.2001047.

[49] K. Seaborn and J. Urakami, ''Measuring voice UX quantitatively: A rapid review,'' in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–8, doi: 10.1145/3411763.3451712.

[50] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, ''Voice in human-agent interaction: A survey,'' *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–43, Jul. 2021, doi: 10.1145/3386867.

[51] G. Caldarini, S. Jaf, and K. McGarry, ''A literature survey of recent advances in chatbots,'' *Information*, vol. 13, no. 1, p. 41, Jan. 2022, doi: 10.3390/info13010041.

[52] A. Arnold, S. Kolody, A. Comeau, and A. M. Cruz, ''What does the literature say about the use of personal voice assistants in older adults? A scoping review,'' *Disability Rehabil., Assistive Technol.*, vol. 19, pp. 1–12, Apr. 2022, doi: 10.1080/17483107.2022.2065369.

[53] R. Gubareva and R. P. Lopes, ''Virtual assistants for learning: A systematic literature review,'' in *Proc. CSEDU*, 2020, pp. 97–103, doi: 10.5220/0009417600970103.

[54] S. Singh and H. Beniwal, ''A survey on near-human conversational agents,'' *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 8852–8866, Nov. 2022, doi: 10.1016/j.jksuci.2021.10.013.

[55] M. Li and A. Suh, ''Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology,'' in *Proc. 54th Hawaii Int. Conf. Syst. Sci.*, 2021, pp. 4053–4062, doi: 10.24251/HICSS.2021.493.

[56] M. Almalki and F. Azeez, ''Health chatbots for fighting COVID-19: A scoping review,'' *Acta Inf. Medica*, vol. 28, no. 4, p. 241, Dec. 2020, doi: 10.5455/aim.2020.28.241-247.

[57] W. Bandara, E. Furtmueller, E. Gorbacheva, S. Miskon, and J. Beekhuyzen, ''Achieving rigor in literature reviews: Insights from qualitative data analysis and tool-support,'' *Commun. Assoc. Inf. Syst.*, vol. 37, pp. 154–204, Jan. 2015, doi: 10.17705/1cais.03708.

[58] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, ''The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration,'' *PLoS Med.*, vol. 6, no. 7, Jul. 2009, Art. no. e1000100, doi: 10.1371/journal.pmed.1000100.

[59] M. Vassar, V. Yerokhin, P. M. Sinnett, M. Weiher, H. Muckelrath, B. Carr, L. Varney, and G. Cook, ''Database selection in systematic reviews: An insight through clinical neurology,'' *Health Inf. Libraries J.*, vol. 34, no. 2, pp. 156–164, Jun. 2017, doi: 10.1111/hir.12176.

[60] S. Carpenter, ''Ten steps in scale development and reporting: A guide for researchers,'' *Commun. Methods Measures*, vol. 12, no. 1, pp. 25–44, Jan. 2018, doi: 10.1080/19312458.2017.1396583.

[61] R. Bernhaupt, D. Drouet, F. Manciet, M. Pirker, and G. Pottier. (2018). *Using Speech To Search: Comparing Built-in and Ambient Speech Search in Terms of Privacy and User Experience*. [Online]. Available: https://www.ibc.org/download?ac=3894

[62] A. Hussein, S. A. Chowdhury, A. Abdelali, N. Dehak, and A. Ali, ''Code-switching text augmentation for multilingual speech processing,'' 2022, *arXiv:2201.02550*.

[63] R. J. De Ayala and M. A. Hertzog, ''The assessment of dimensionality for use in item response theory,'' *Multivariate Behav. Res.*, vol. 26, no. 4, pp. 765–792, Oct. 1991, doi: 10.1207/s15327906mbr2604_9.

[64] R. F. DeVellis, *Scale Development: Theory and Applications*, vol. 26, 3rd ed. Newbury Park, CA, USA: Sage, 2012, p. 31. [Online]. Available: http://books.google.com/books?id=Rye31saVXmAC&lpg=PR1&ots=YHXbaKkzn1&dq=affectivewritingscalevalidityexamples&lr&pg=PR1#v=onepage&q&f=false

[65] E. Anthoine, L. Moret, A. Regnault, V. Sbille, and J. B. Hardouin, ''Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures,'' *Health Quality Life Outcomes*, vol. 12, no. 1, pp. 1–10, Dec. 2014, doi: 10.1186/S12955-014-0176-2.

[66] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, ''Sample size in factor analysis,'' *Psychol. Methods*, vol. 4, no. 1, pp. 84–99, Mar. 1999, doi: 10.1037/1082-989x.4.1.84.

[67] *Tabachnick & Fidell, Using Multivariate Statistics, 6th Edition | Pearson*. Accessed: May 14, 2022. [Online]. Available: https://www.pearson.com/us/higher-education/program/Tabachnick-Using-Multivariate-Statistics-6th-Edition/PGM332849.html

[68] B. Price, ''A first course in factor analysis,'' *Technometrics*, vol. 35, no. 4, p. 453, Nov. 1993, doi: 10.1080/00401706.1993.10485363.

[69] R. L. Worthington and T. A. Whittaker, ''Scale development research: A content analysis and recommendations for best practices,'' *Counseling Psychologist*, vol. 34, no. 6, pp. 806–838, Nov. 2006, doi: 10.1177/0011000006288127.

[70] M. R. Islaqm and M. R. Islam, ''Sample size and its role in Central Limit Theorem (CLT),'' *Comput. Appl. Math. J.*, vol. 4, no. 1, pp. 1–7, 2018. [Online]. Available: http://www.aascit.org/journal/camj

[71] H. T. Schreuder, T. G. Gregoire, and J. P. Weyer, ''For what applications can probability and non-probability sampling be used?'' *Environ. Monitor. Assessment*, vol. 66, no. 3, pp. 281–291, 2001, doi: 10.1023/a:1006316418865.

[72] R. L. Gorsuch, ''Exploratory factor analysis: Its role in item analysis,'' *J. Personality Assessment*, vol. 68, no. 3, pp. 532–560, Jun. 1997, doi: 10.1207/s15327752jpa6803_5.

[73] M. Auerswald and M. Moshagen, ''How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions,'' *Psychol. Methods*, vol. 24, no. 4, pp. 468–491, Aug. 2019, doi: 10.1037/met0000200.

[74] D. Harrington, *Confirmatory Factor Analysis*. U.K.: Oxford Univ. Press, Jan. 2009, pp. 1–128, doi: 10.1093/ACPROF:OSO/9780195339888.001.0001.

[75] D. Hooper, J. Coughlan, and M. R. Mullen, ''Structural equation modelling: Guidelines for determining model fit,'' *Electron. J. Bus. Res. Methods*, vol. 6, no. 1, pp. 53–60, 2008.

[76] R. H. Hoyle, ''Confirmatory factor analysis,'' in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. New York, NY, USA: Academic, 2000, pp. 465–497, doi: 10.1016/B978-012691360-6/50017-3.

[77] L. L. Chan and N. Idris, ''Validity and reliability of the instrument using exploratory factor analysis and Cronbach's alpha,'' *Int. J. Academic Res. Bus. Social Sci.*, vol. 7, no. 10, p. 400, Oct. 2017, doi: 10.6007/ijarbss/v7-i10/3387.

[78] P. F. M. Krabbe, ''Validity,'' in *The Measurement of Health and Health Status*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 113–134, doi: 10.1016/B978-0-12-801504-9.00007-6.

[79] J. H. Amirkhan, ''Criterion validity of a coping measure,'' *J. Personality Assessment*, vol. 62, no. 2, pp. 242–261, 2010, doi: 10.1207/S15327752JPA6202_6.

[80] D. Granpeesheh, J. Tarbox, A. C. Najdowski, J. Kornack, J. Tarbox, and A. C. Najdowski, ''Chapter 19—Clinical supervision,'' in *Evidence-Based Treatment for Children With Autism: The CARD Model*. Amsterdam, The Netherlands: Elsevier, 2014, doi: 10.1016/B978-0-12-411603-0.00019-7.

[81] L. B. D. Gottems, E. M. P. D. Carvalho, D. Guilhem, and M. R. M. Pires, ''Good practices in normal childbirth: Reliability analysis of an instrument by Cronbach's alpha,'' *Revista Latino-Americana de Enfermagem*, vol. 26, p. 3000, May 2018, doi: 10.1590/1518-8345.2234.3000.

[82] K. S. Hone and R. Graham, ''Towards a tool for the subjective assessment of speech system interfaces (SASSI),'' *Natural Lang. Eng.*, vol. 6, pp. 287–303, Sep. 2000, doi: 10.1017/s1351324900002497.

[83] L. Wulf, M. Garschall, J. Himmelsbach, and M. Tscheligi, ''Hands free-care free: Elderly people taking advantage of speech-only interaction,'' in *Proc. 8th Nordic Conf. Human-Comput. Interact., Fun, Fast, Foundational*, 2014, pp. 203–206, doi: 10.1145/2639189.2639251.

[84] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and ergonomic quality aspects determine a software's appeal," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2000, pp. 201–208, doi: 10.1145/332040.332432.

[85] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness," *Int. J. Hum.-Comput. Interact.*, vol. 13, no. 4, pp. 481–499, Dec. 2001, doi: 10.1207/s15327590ijhc1304_07.

[86] J. R. Lewis. (2001). *Psychometric Properties of the Mean Opinion Scale Usability Evaluation and Interface Design-Proceedings of HCI International 2001 (Mahwah, NJ: Lawrence Erlbaum) Psychometric Properties of the Mean Opinion Scale*. [Online]. Available: https://www.researchgate.net/publication/252235259

[87] J. Cambre, J. Colnago, and J. Tsai, "Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[88] M. D. Polkosky and J. R. Lewis, "Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X," *Int. J. Speech Technol.*, vol. 6, no. 2, pp. 161–182, Apr. 2003, doi: 10.1023/a:1022390615396.

[89] P. S. Polatoğlu. (2019). *Gender and Style Effects in Voice Assistants*. [Online]. Available: https://theses.liacs.nl/pdf/2018-2019-PolatogluPS.pdf

[90] M. Polkosky. (Feb. 2005). *Toward a Social-Cognitive Psychology of Speech Technology: Affective Responses To Speech-Based E-Service*. Accessed: Apr. 21, 2022. [Online]. Available: https://digitalcommons.usf.edu/etd/819

[91] M. Amith, R. Lin, R. Cunningham, Q. L. Wu, L. S. Savas, Y. Gong, J. A. Boom, L. Tang, and C. Tao, "Examining potential usability and health beliefs among young adults using a conversational agent for HPV vaccine counseling," *AMIA Summits Transl. Sci. Proc.*, vol. 2020, pp. 43–52, Jan. 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32477622%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7233050

[92] J. R. Lewis and M. L. Hardzinski, "Investigating the psychometric properties of the speech user interface service quality questionnaire," *Int. J. Speech Technol.*, vol. 18, no. 3, pp. 479–487, Sep. 2015, doi: 10.1007/s10772-015-9289-1.

[93] F. Iniesto, T. Coughlan, K. Lister, P. Devine, N. Freear, R. Greenwood, W. Holmes, I. Kenny, K. McLeod, and R. Tudor, "Creating 'a simple conversation': Designing a conversational user interface to improve the experience of accessing support for study," *ACM Trans. Accessible Comput.*, vol. 16, no. 1, pp. 1–29, Mar. 2023, doi: 10.1145/3568166.

[94] H. Jung, H. Kim, and J.-W. Ha, "Understanding differences between heavy users and light users in difficulties with voice user interfaces," in *Proc. 2nd Conf. Conversational User Interfaces*, Jul. 2020, pp. 1–4, doi: 10.1145/3405755.3406170.

[95] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Proc. Symp. Austrian HCI Usability Eng. Group* (Lecture Notes in Computer Science), vol. 5298, 2008, pp. 63–76, doi: 10.1007/978-3-540-89350-9_6.

[96] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 4, no. 6, p. 103, 2017, doi: 10.9781/ijimai.2017.09.001.

[97] T. Kannampallil, C. R. Ronneberg, N. E. Wittels, V. Kumar, N. Lv, J. M. Smyth, B. S. Gerber, E. A. Kringle, J. A. Johnson, P. Yu, L. E. Steinman, O. A. Ajilore, and J. Ma, "Design and formative evaluation of a virtual voice-based coach for problem-solving treatment: Observational study," *JMIR Formative Res.*, vol. 6, no. 8, Aug. 2022, Art. no. e38092, doi: 10.2196/38092.

[98] A. M. Klein, A. Hinderks, M. Schrepp, and J. Thomaschewski. *Construction of UEQ+ Scales for Voice Quality Measuring User Experience Quality of Voice Interaction*. [Online]. Available: https://ueqplus.ueq-research.org/

[99] G. Haas, M. Rietzler, M. Jones, and E. Rukzio, "Keep it short: A comparison of voice assistants' response behavior," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2022, pp. 1–12, doi: 10.1145/3491102.3517684.

[100] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design* (Lecture Notes in Computer Science), vol. 5619. New York, NY, USA: Springer, 2009, pp. 94–103, doi: 10.1007/978-3-642-02806-9_12.

[101] M. Turunen, J. Hakulinen, A. Melto, T. Heimonen, T. Laivo, and J. Hella, "SUXES—User experience evaluation method for spoken and multimodal interaction," in *Proc. INTERSPEECH*, 2009, pp. 2567–2570, doi: 10.21437/INTERSPEECH.2009-676.

[102] M. Turunen, J. Hakulinen, A. Melto, and J. Hella, "Speech-based and multimodal media center for different user groups speech-based and multimodal media center for different user groups," in *Proc. 10th Annu. Conf. Int. Speech Commun. Association*, Sep. 2009, pp. 1–4, doi: 10.21437/Interspeech.2009-441.

[103] I. Wechsung, B. Weiss, C. Kühnel, P. Ehrenbrink, and S. Möller, "Development and validation of the conversational agents scale (CAS)," in *Proc. Interspeech*, Aug. 2013, pp. 1106–1110, doi: 10.21437/interspeech.2013-298.

[104] U. Saad, U. Afzal, A. El-Issawi, and M. Eid, "A model to measure QoE for virtual personal assistant," *Multimedia Tools Appl.*, vol. 76, no. 10, pp. 12517–12537, May 2017, doi: 10.1007/s11042-016-3650-5.

[105] K. Israfilzade, "Conversational marketing as a framework for interaction with the customer: Development & validation of the conversational agent's usage scale," *J. LIFE Econ.*, vol. 8, no. 4, pp. 533–546, Oct. 2021, doi: 10.15637/jlecon.8.4.12.

[106] G. Guerino, W. Silva, T. Coleti, and N. Valentim, "Assessing a technology for usability and user experience evaluation of conversational systems: An exploratory study," in *Proc. 23rd Int. Conf. Enterprise Inf. Syst.*, Apr. 2021, pp. 463–473, doi: 10.5220/0010450204630473.

[107] L. I. D. Faruk, S. Funilkul, P. Mongkolnam, P. Puengwattanapong, and D. Pal, "Exploring user experience with voice assistants: Impact of prior experience on voice assistants," in *Proc. 13th Int. Conf. Adv. Inf. Technol.*, Dec. 2023, pp. 1–9, doi: 10.1145/3628454.3629470.

[108] R. Phinnemore, M. Reza, B. Lewis, K. Mahadevan, B. Wang, M. Annett, and D. Wigdor, "Creepy assistant: Development and validation of a scale to measure the perceived creepiness of voice assistants," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2023, pp. 1–18, doi: 10.1145/3544548.3581346.

[109] J. M. Lee, J. Baek, and D. Y. Ju, "Anthropomorphic design: Emotional perception for deformable object," *Frontiers Psychol.*, vol. 9, p. 1829, Oct. 2018, doi: 10.3389/fpsyg.2018.01829.

[110] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Experim. Social Psychol.*, vol. 52, pp. 113–117, May 2014, doi: 10.1016/j.jesp.2014.01.005.

[111] C. Nass, I. M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2005, pp. 1973–1976, doi: 10.1145/1056808.1057070.

[112] D. G. School and M. D. Polkosky. (2005). *Scholar Commons Toward a Social-Cognitive Psychology of Speech Technology: Affective Responses To Speech-Based E-Service*. [Online]. Available: https://scholarcommons.usf.edu/etd

[113] K. Catherine. (2022). *AI Assistant Tech Increasingly Popular in Australia*. Accessed: Sep. 21, 2022. [Online]. Available: https://itbrief.com.au/story/ai-assistant-tech-increasingly-popular-in-australia

[114] Carlos Mureithi. (2021). *Siri and Alexa Still Don't Support African Languages—Quartz Africa*. Accessed: Sep. 21, 2022. [Online]. Available: https://qz.com/africa/2014030/siri-and-alexa-still-dont-support-african-languages/

[115] H. B. Santoso and M. Schrepp, "The impact of culture and product on the subjective importance of user experience aspects," *Heliyon*, vol. 5, no. 9, Sep. 2019, Art. no. e02434, doi: 10.1016/j.heliyon.2019.e02434.

[116] T. R. B. de Souza and J. L. Bernardes, "The influences of culture on user experience," in *Proc. Int. Conf. Cross-Cultural Design* (Lecture Notes in Computer Science), vol. 9741, 2016, pp. 43–52, doi: 10.1007/978-3-319-40093-8_5.

[117] D. Pal, C. Arpnikanondt, S. Funilkul, and V. Varadarajan, "User experience with smart voice assistants: The accent perspective," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–6, doi: 10.1109/ICCCNT45670.2019.8944754.

[118] N. Bevan, "Usability is quality of use," *Adv. Hum. Factors/Ergonom.*, vol. 20, pp. 349–354, Jan. 1995, doi: 10.1016/S0921-2647(06)80241-8.

[119] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*, vol. 346. VI, USA: Springer, 2011, pp. 623–654, doi: 10.1007/978-3-642-19551-8_23.

[120] S. Reysen, "Construction of a new scale: The Reysen likability scale," *Social Behav. Personality, Int. J.*, vol. 33, no. 2, pp. 201–208, Jan. 2005, doi: 10.2224/sbp.2005.33.2.201.

[121] A. Finn and U. Kayande, "Scale modification: Alternative approaches and their consequences," *J. Retailing*, vol. 80, no. 1, pp. 37–52, Jan. 2004, doi: 10.1016/j.jretai.2004.01.003.

[122] M. Braun, A. Mainz, R. Chadowitz, B. Pfleging, and F. Alt, "At your service," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–11, doi: 10.1145/3290605.3300270.

[123] J. Kengphanich, J. Buatip, and S. Phithakkitnukoon, "Eventity: Online platform for city event and tourism information," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, Oct. 2018, pp. 1051–1059, doi: 10.1145/3267305.3274160.

[124] P. Prommaharaj, S. Phithakkitnukoon, M. Veloso, and C. Bento, "Visualization tool for taxi usage analysis: A case study of Lisbon, Portugal," in *Proc. ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Adjun.*, 2016, pp. 1343–1348, doi: 10.1145/2968219.

[125] C. Somdulyawat, P. Pongjitpak, S. Phithakkitnukoon, M. Veloso, and C. Bento, "A tool for exploratory visualization of bus mobility and ridership: A case study of Lisbon, Portugal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, 2015, pp. 1117–1121, doi: 10.1145/2800835.2800975.

[126] M. B. Yilmaz and K. Rizvanoglu, "Understanding users' behavioral intention to use voice assistants on smartphones through the integrated model of user satisfaction and technology acceptance: A survey approach," *J. Eng., Design Technol.*, vol. 20, no. 6, pp. 1738–1764, Dec. 2022, doi: 10.1108/jedt-02-2021-0084.

[127] R. Zhong, M. Ma, Y. Zhou, Q. Lin, L. Li, and N. Zhang, "User acceptance of smart home voice assistant: A comparison among younger, middle-aged, and older adults," *Univers. Access Inf. Soc.*, pp. 1–18, Oct. 2022, doi: 10.1007/s10209-022-00936-1.

[128] L. I. D. Faruk, R. Rohan, U. Ninrutsirikun, and D. Pal, "University students' acceptance and usage of generative AI (ChatGPT) from a psycho-technical perspective," in *Proc. 13th Int. Conf. Adv. Inf. Technol.*, Dec. 2023, pp. 1–8, doi: 10.1145/3628454.3629552.

[129] S. Borsci, M. Schmettow, A. Malizia, A. Chamberlain, and F. van der Velde, "A confirmatory factorial analysis of the chatbot usability scale: A multilanguage validation," *Pers. Ubiquitous Comput.*, vol. 27, no. 2, pp. 317–330, Apr. 2023, doi: 10.1007/s00779-022-01690-0.

LAWAL IBRAHIM DUTSINMA FARUK received the B.Sc. degree from the University of Portsmouth, England, and the master's degree from Mae Fah Luang University, Thailand. He is currently pursuing the Ph.D. degree with the King Mongkut's University of Technology Thonburi. He is focused on researching user experience with artificial intelligent voice agents and dialogue flow capabilities for his doctoral work. His passion extends beyond academia, aiming to bridge the gap between emergent technology and human needs, particularly in the realm of assistive technologies. His research interests include human–computer interaction, user experience, assistive emergent technology, and AI tools.

MOHAMMAD DAWOOD BABAKERKHELL was born in Khost, Afghanistan. He received the B.Sc. degree in computer science from Shaikh Zayed University (Khost), Afghanistan, in 2009, and the M.Sc. degree in network technology and management from the Amity Institute of Information Technology, Amity University, Uttar Pradesh, India, in 2019. He has been a Lecturer with the Information Technology Department, Computer Science Faculty, Shaikh Zayed University (Khost), since 2011. His research interests include information technology management, computer network and security, cloud computing, and the Internet of Things.

PORNCHAI MONGKOLNAM received the Ph.D. degree in computer science from Arizona State University, with a focus on software engineering, artificial neural networks, and artificial intelligence. He is currently an Associate Professor with the School of Information Technology, King Mongkut's University of Technology Thonburi. His research interests include innovative research on stress and mood recognition systems and practical applications, like comparing and normalizing the measurement of step counts and heart rates of selected wristbands and smartwatches.

VITHIDA CHONGSUPHAJAISIDDHI received the M.A. degree in media technology for TEFL from the University of Newcastle upon Tyne, U.K., in 1996, and the M.Sc. degree in computing science and the Ph.D. degree in information technology from the King Mongkut's University of Technology Thonburi, in 1997. She is currently a Lecturer with the School of Information Technology and the Chair of the Bachelor of Arts Program in Digital Service Innovation, King Mongkut's University of Technology Thonburi. Her research interests include human–computer interaction and software engineering.

SUREE FUNILKUL received the B.Sc. degree in science (mathematics) from Mahidol University, in 1996, and the M.Sc. and Ph.D. degrees in information technology from the King Mongkut's University of Technology Thonburi, in 2000 and 2008, respectively. She is currently a Lecturer with the King Mongkut's University of Technology Thonburi. Her impactful research, "E-Democracy System Development Framework and Its Quality," underscores her commitment to technological advancement for societal benefit. Her research interests include information systems and database programming.

DEBAJYOTI PAL received the B.E. degree in electrical engineering from Nagpur University, Maharashtra, India, in 2005, the M.Tech. degree in information technology from the Indian Institute of Engineering Science and Technology, Shibpur, Kolkata, India, in 2007, and the Ph.D. degree in information technology from the School of IT, KMUTT, Bangkok, Thailand. He is currently a Lecturer with KMUTT. His research interests include technology acceptance, the IoT, human–computer interaction, QoS/QoE, user experience design, ambient intelligence, and multimedia service evaluation.