Understanding How to Administer Voice Surveys through Smart Speakers

JING WEI, University of Melbourne, Australia
WEIWEI JIANG, University of Melbourne, Australia
CHAOFAN WANG, University of Melbourne, Australia
DIFENG YU, University of Melbourne, Australia
JORGE GONCALVES, University of Melbourne, Australia
TILMAN DINGLER, University of Melbourne, Australia
VASSILIS KOSTAKOS, University of Melbourne, Australia

Smart speakers have become exceedingly popular and entered many people's homes due to their ability to engage users with natural conversations. Researchers have also looked into using smart speakers as an interface to collect self-reported health data through conversations. Responding to surveys prompted by smart speakers requires users to listen to questions and answer in voice without any visual stimuli. Compared to traditional web-based surveys, where users can see questions and answers visually, voice surveys may be more cognitively challenging. Therefore, to collect reliable survey data, it is important to understand what types of questions are suitable to be administered by smart speakers. We selected five common survey questionnaires and deployed them as voice surveys and web surveys in a within-subject study. Our 24 participants answered questions using voice and web questionnaires in one session. They then repeated the same study session after 1 week to provide a "retest" response. Our results suggest that voice surveys have comparable reliability to web surveys. We find that, when using 5-point or 7-point scales, voice surveys take about twice as long as web surveys. Based on objective measurements, such as response agreement and test-retest reliability, and subjective evaluations of user experience, we recommend that researchers consider adopting the binary scale and 5-point numerical scales for voice surveys on smart speakers.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Sound-based input / output**; **Empirical studies in interaction design**.

Additional Key Words and Phrases: smart speakers, voice user interfaces, survey methodology, chatbots, conversational user interfaces

ACM Reference Format:
Jing Wei, Weiwei Jiang, Chaofan Wang, Difeng Yu, Jorge Goncalves, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding How to Administer Voice Surveys through Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 548 (November 2022), 32 pages. https://doi.org/10.1145/3555606

Authors' addresses: Jing Wei, jing.wei@student.unimelb.edu.au, University of Melbourne, Melbourne, Australia; Weiwei Jiang, weiwei.jiang@student.unimelb.edu.au, University of Melbourne, Melbourne, Australia; Chaofan Wang, chaofanw@student.unimelb.edu.au, University of Melbourne, Melbourne, Australia; Difeng Yu, difeng.yu@student.unimelb.edu.au, University of Melbourne, Melbourne, Australia; Jorge Goncalves, jorge.goncalves@unimelb.edu.au, University of Melbourne, Melbourne, Australia; Tilman Dingler, tilman.dingler@unimelb.edu.au, University of Melbourne, Melbourne, Australia; Vassilis Kostakos, vassilis.kostakos@unimelb.edu.au, University of Melbourne, Melbourne, Australia.

2573-0142/2022/11-ART548 $15.00
https://doi.org/10.1145/3555606

1 INTRODUCTION

The popularity of smart speakers is soaring [31]. The functionality of smart speakers has extended well beyond basic music playing or smart appliance controlling. Researchers have investigated how smart speakers can be used to educate children [21], accompany older adults [59], and provide mental support for those in needs [62, 83]. Moreover, other than providing information to users, smart speakers are poised to become a practical tool for collecting user data via conversations [20, 25, 42, 61]. For example, Luo et al. [42] have developed an Alexa skill that allows users to log their exercise data through hands-free and eyes-free conversations. Maharjan et al. [45] used smart speakers to administer a 5-item health-related questionnaire and suggested that smart speakers could be a viable platform to administer surveys and support health and wellbeings.

In the Computer-Supported Cooperative Work (CSCW) and Human-Computer Interaction (HCI) research community, surveys are widely used to acquire self-reported opinions from the crowd [11], gather data for intervention systems [82], collect frequent ESM self-reports of user sociometrics, such as creativity [79], mood, and feelings [46, 47]. We envision that smart speakers can offer a new modality for researchers or other stakeholders to reach users and collect data, either opinions or personal self-reports, through voice surveys. However, it is unclear whether data from such voice surveys can be as reliable and valid as pen&paper surveys or web/mobile surveys [27].

The widely used Interactive Voice Response (IVR) surveys [19] are similar to smart speaker-enabled voice surveys where survey questions are delivered via audio to respondents. To fill in traditional IVR surveys, respondents need to press numbers on a keypad [35]. More recent work has implemented IVR surveys with voice inputs using modern Automatic Speech Recognition (ASR) technologies to reach people from low-income and less-literate populations [29, 63]. The popularity of IVR surveys inspired us to investigate whether smart speakers can be an emerging platform to administer such surveys. Smart speakers can augment traditional IVRs by taking into consideration the owner's personal preferences, such as the gender of the voice or accessibility features. Voice surveys on smart speakers have the potential to be more personalized than IVR surveys.

The way smart speakers interact with users can also impact the administration of voice surveys [35]. Smart speakers are accessible and accept natural voice inputs. Hence, compared to typing wordy answers on a keypad, respondents may be more willing to give thoughtful answers to open-ended questions on smart speakers [64]. Close-ended questions, such as 5-point and 7-point Likert scales, are commonly used to measure levels of agreement or user satisfaction in CSCW [11]. The 7-point scale is considered to be the optimal scale in most visually-administered surveys for its excellent validity; however, there is still uncertainty about the usability of the 7-point in telephone surveys [35]. The main drawback of smart speakers is that they provide minimal visual feedback. Without visual references, respondents may experience a higher cognitive burden when responding with longer lengths scales through voice [42, 45]. Recording answers through speech is also more error-prone than using keypads. ASR in smart speakers is far from perfect, and users sometimes need to repeat their speech multiple times [52, 58]. Furthermore, interacting with smart speakers can sometimes be slow as VUIs do not allow quick "error recovery": in the case where respondents need to re-hear the question, they need to utter specific commands to request the speaker to repeat or even restart. Extended survey completion time negatively impacts respondents' motivation [7], especially for ESM-style surveys [81]. These limitations make voice surveys on smart speakers difficult to respond to, which may lead to compromised data quality [26, 37] and user reluctance to adopt this survey modality.

Motivated by the promising use of smart speakers for survey administrations and the potential challenges of cognitive demands and error-prone interaction with voice surveys, our work aims to address one overarching research question: *What types of questions are suitable to be deployed in*

voice surveys administered by smart speakers? We selected five commonly used question types to be evaluated: open-ended questions, binary questions (yes/no), 5-point Likert scale questions answered with numbers (we use the term *5-point #* in the following), 5-point Likert scale questions answered with labels (e.g., agree/disagree, we use the term *5-point W* in the following), and 7-point scale questions answered with numbers (we use the term *7-point #* in the following). We implemented five survey questionnaires as voice surveys on Google speakers and those surveys' counterparts as web surveys on Qualtrics. In a within-subjects study, 24 participants responded to all voice and web surveys and evaluated different types of survey questions with subjective quantitative forms and qualitative interview feedback. Our study examines the "suitability" of different types of questions based on three aspects: 1) the validity and reliability of questions, 2) user perceptions and preferences, and 3) the recognition error rates of different formats of answer options.

Our work makes three key contributions to future CSCW research that implements and investigates voice surveys on smart speakers: first, we empirically show that these voice surveys overall have good reliability except for the 5-point W scale, which performs slightly worse in terms of reliability and validity. The user subjective evaluation and interview feedback also confirm that most participants consider 5-point W more difficult to respond to. Second, we explore the user perception and experience of smart speakers as a new survey modality. We show that although participants believe voice surveys are more difficult to use and are error-prone, they also consider voice surveys to be more fun to use and have the potential to allow multitasking in many scenarios. Third, we elicit recommendations for rating scales and discuss implications for designing voice surveys on smart speakers. We also point out how future research can build on our findings to examine voice surveys.

2 RELATED WORK

2.1 Modality Differences

Commonly used survey modalities include face-to-face, telephone (e.g., IVR systems), mail, web, or mixed-mode surveys [19]. Previous studies suggest that the modality of surveys can impact the quality of responses [19, 72]. For example, adding aurally-spoken questions to visually based questionnaires is found to increase the report of "stigmatized behaviors" and induce more self-disclosure [15, 80]. Compared to traditional mail-in or face-to-face surveys, IVR surveys are found to yield more positive and extreme answers [18]. More recently, researchers have implemented questionnaires using chatbots and made the responding process like a conversation [30, 88]. The humanization nature of conversational interfaces appears to improve the enjoyability [8], the level of self-disclosure [88], and the quality of response data [30].

Since both chatbot surveys and A-CASI are found to improve the response quality, we wonder whether smart speakers, which implement voice-based conversations, can also be a potential platform to administer surveys. However, voice surveys administered by smart speakers are much different than chatbot or telephone surveys. Chatbot surveys provide conversational interactions, but they are deployed on Graphical User Interfaces (GUIs). Respondents can visually see questions and options. For close-ended questions, they can choose answers by clicking. To respond to keypad-based IVR surveys, respondents hear all the questions aurally and quickly respond with keypads, which also provide some visual references. Compared to those modalities, the major drawback of voice surveys on smart speakers is the lack of visual representation of answer options. Respondents need to think of answers in their minds, which might be challenging. The limited speech recognition algorithms increase the survey difficulty – respondents' answers may not be successfully recognized every time.

2.2 Survey Questionnaire Design

The design of survey questionnaires has been found to impact the reliability and validity of responses. To give an optimal answer, it is commonly agreed that respondents need to understand the question (comprehension), search their memories for relevant information (retrieval), make judgment based on the information (judgment), and finally translate their judgment into a response (response selection/matching) [35, 78]. These four steps, however, can be cognitively challenging. When facing a great amount of cognitive demands from surveys, some people may use satisficing to reduce the burden, i.e., they provide satisfactory answers rather than optimal answers [34]. In particular, respondents are more likely to satisfice if the questionnaire design is poor, e.g., questions are difficult to interpret, or the response selection is somewhat complicated (e.g., "10" in a keypad is less chosen as pressing it is more complicated than pressing "9"). Satisficing can cause respondents to think superficially and give answers they consider acceptable rather than think thoroughly and give genuine answers. Krosnick [34] identified three factors that are correlated to satisficing: *task difficulty*, *respondent ability*, *respondent motivation*. Among the three factors, researchers can control the *task difficulty* of questionnaires to discourage satisficing [35].

In particular, task difficulty is determined by question-specific attributes and the administration of questionnaires [35]. Compared to close-ended questions, open-ended questions allow respondents to use their own words and discourage satisficing as they eliminate the response selection step [55]. However, responses to open-ended questions require more effort to interpret and analyze. On the other hand, close-ended questions, where respondents are given a set of options to choose from [67], are easier to administer. In the CSCW community, the 5-point scale is widely used [41]. Other commonly used scales include the 2-point binary scale (yes/no or true/false) [3], the 7-point scale [53], and even the 11-point scale [11]. To design optimal close-ended questions, it is important to consider the points of rating scales [14, 55]. Existing literature suggests that moderately long scale lengths (i.e., more points) can increase both the reliability [23, 75] and the validity [68] of questionnaires. In particular, Krosnick [35] suggest that the 7-point scale is considered optimal for most visually administered questionnaires. But for traditional IVR surveys, it is still unclear whether the 7-point is the optimal scale.

As aforementioned, voice surveys may be intrinsically more difficult due to the nature of VUI. The voice question design is therefore of great importance. If open-ended questions are administered through smart speakers, respondents may encounter interaction errors [52] and the recorded answers may also contain transcription errors [28, 84]. On the other hand, since speaking is faster and easier than typing, the voice may be more advantageous than the web in accepting free-form longer answers. For close-ended questions, due to the lack of visual presentation of rating scales, longer rating scales may not work well in voice surveys. The binary scale lacks refinement in choices [3], but it may work well in voice surveys due to its simplicity. For commonly used 5-point scales, previous work suggests that scales labeled with words (e.g., agree) can achieve higher reliability than those labeled with numbers (e.g., 4 - agree) as word labels are closer to respondents' mental representation of answers and reduce the effort in matching respondents' thinking to numbers [35]. However, while the answer "strongly disagree" may be easier to think of, uttering it is literally more difficult than speaking "1". So, it remains unclear whether a numerical scale (i.e., respondents answer with numbers from 1 to 5) or a word scale should be used in voice surveys. Lastly, while the 7-point scale is considered "optimal" for some survey questionnaires, it may be too hard to use for voice surveys. Drawn on prior work on questionnaire design, we are interested in investigating different voice survey questionnaire designs on smart speakers.

2.3 IVR-based Questionnaires and Surveys

Many IVR surveys allow respondents to hear spoken questions from their phones and provide answers by pressing the keypad [18]. Prior work in developing countries has suggested that IVR voice surveys can include a short, straightforward introductory message of the survey to improve the completion rate, and use shorter question prompts (without response options) to improve data quality [43, 76]. Researchers have also investigated using voice inputs for IVR surveys. Before the recent advances in ASR technologies [16], a study dating back to 1989 implemented IVR surveys that allow voice inputs [13]. With the use of traditional voice recognition algorithms, Clayton and Winter [13] suggested that the voice survey had a high response rate and was an alternative way to collect data. They also reported that some respondents preferred the voice inputs to the keypad inputs and considered it was fast to use voice inputs. Similarly, another early work in 2005 found that most users preferred the voice inputs to keypad inputs despite encountering fairly high recognition errors [39]. With today's advanced speech recognition technologies, Khullar et al. [29] implemented IVR surveys that have various question types: multiple choice questions (MCQ) and binary questions. They showed that natural voice-based surveys had great accuracy and had comparable performance to keypad-based surveys, and their participants had a strong preference for the voice inputs. They also pointed out that numerical questions were easier to answer through voice. Randhawa et al. [63] implemented a voice-based crowdsourcing platform called Karamad, which supports binary questions, MCQ and open-ended questions. Interestingly, they found that their participants answered more consistently to binary questions and suggested that a lower consistency in MCQ may be due to the increased cognitive load in text-free tasks. These works, together with previous studies, have shown that respondents from low-literate populations and developing regions benefit from voice-based IVR surveys [56, 71]. In fact, IVR surveys are mostly advantageous as they can be easily distributed in regions where the access to smartphones and stable internet is limited, through regular phone calls [10, 40].

While much research has focused on IVR surveys, it is unclear whether design guidelines for IVR surveys can be generalized to voice surveys on smart speakers. There is a need to explore which question types, for example, binary questions or MCQ, should be used on smart speakers [63]. Like voice-based IVR surveys, surveys delivered by smart speakers may also introduce a high cognitive load that could hamper response quality [32, 63]. Recent studies have further suggested that interactions with speakers can be error-prone and subject to network delays [58, 85]. Such challenges can especially hinder the successful adaption of voice surveys in the context of digital health systems. Regular check-ins with patients and the collection of self-reports (e.g., medical or mental condition and activities) require robust interactions [11, 20, 42]. It is thus essential to evaluate the validity and reliability of different question types on smart speakers.

2.4 Conversational Agent-based Questionnaires and Surveys

With the increasing popularity of chatbots, there is a plethora of research that focuses on conversational surveys. Xiao et al. [88] implemented a chatbot that asked open-ended questions and compared the chatbot survey with the traditional web survey on Qualtrics. They analyzed the free-text responses and suggested that the chatbot induced a higher level of engagement and elicited better quality responses. Similarly, Celino and Calegari [8] showed that submitting responses to conversational surveys can achieve the same reliability and a higher response quality with respect to traditional web surveys. Kim et al. [30] implemented both chatbot surveys and web surveys that administer mostly close-ended questions. They used the satisficing theory to compare the responses' data quality from two modalities and found that participants provided more differentiated responses and satisfice less when responding to the chatbot survey. The authors also suggested

that the chatbot with a casual conversation style can create more user engagement. Further, Rhim et al. [65] compared a baseline chatbot with a humanized chatbot that used humanizing traits like self-introduction, adaptive response speed, and echoing. They found that users perceived the humanized survey chatbot more positively and reported to have higher level satisfaction levels and more self-disclosure in their responses. Compared to conversational text-based chatbots, voice assistants are more "human-like" and engage users with natural voices. Enabling voice-based conversational surveys may further increase user engagement and discourage satisficing.

Recently, smart speakers have been used to collect data through conversations. For example, voice-based Experience Sampling Method (ESM) applications have been developed and deployed on smart speakers [9, 84]. Also, self-tracking systems have been developed to allow users to report health and wellness data by talking to smart speakers [42]. Researchers have also looked into implementing standard survey questionnaires on smart speakers. Maharjan et al. [45] transformed an established 5-item questionnaire to be a voice survey on Google speakers. They evaluated two response inputs, discrete and open-ended, to the voice survey and found that users preferred the discrete response inputs. They also compared verbal responses with paper-based responses and suggested that smart speakers can be a feasible platform for collecting data through voice surveys. However, as they only used one short questionnaire, the questionnaire difficulty was relatively low, and the participants could easily remember their paper-based responses. Their results cannot be extended to inform the design of other questionnaires (e.g., binary and 7-point).

Previous works on voice-based IVR systems have shed light on the use of voice surveys and provided some evidence on the reliability of MCQ questions and binary questions in voice surveys [63]. Works on chatbots have also shown that conversation can improve data quality. However, smart speakers interact with people differently, and they are envisioned to measure people's feelings, thoughts, and psychometrics [44, 81] with standard questionnaires [45] or Likert scales [84]. It is important to ensure that users feel comfortable talking to smart speakers and provide reliable self-reports or opinions. However, to the best of our knowledge, it is still unknown whether screenless smart speakers can administer conversational voice surveys and collect comparable or even better quality data than vision-based surveys. Hence, we want to examine the validity and reliability of five types of questions in voice surveys in comparison to the commonly used web surveys and investigate how people respond to and perceive those questions in this study.

3 METHODS

3.1 Study Design

3.1.1 Questionnaire Selection and Survey Design. The main goal of our study is to investigate what types of questions are suitable to be deployed on smart speakers. In the CSCW community, there are a few commonly used question types: Likert scales of different lengths, dichotomous scales, and open-ended questions. To investigate whether each form of questions can be delivered and answered in voice, we select four established survey questionnaires to be used in our study: 1) Positive Thinking Scale (PTS, 22 yes/no items) [17], 2) The Behavioural Regulation In Exercise Questionnaire-3 (BREQ-3, 24 5-point scale items) [48, 87], 3) Problematic Use of Mobile Phones Scale (PUMP, 20 5-point scale items) [50], and 4) Fear of Spiders Questionnaire (FSQ, 18 7-point scale items) [77]. Two 5-point scale surveys are chosen as we want to compare the use of numerical scales and word scales. We make the BREQ-3 to be answered with numerical labels (i.e., "1" to "5") and the PUMP scale to be answered with word labels (i.e., "strongly disagree" to "strongly agree"). For the 7-point scale, we do not implement the word scale form as we think it is too troublesome to remember 7 word expressions [36]. We also design an 11-item demographic questionnaire (DQ) that asks about people's age, gender, education, and experiences with smart speakers in an open-ended

Table 1. The allocation of items for the four selected questionnaires. The Binary scale (PTS) has 11 positive items (P) and 11 negative items (N). The 5-point # scale (BREQ-3) has 6 dimensions (dim.), and every dimension is measured with 4 items. The 5-point W scale (PUMP) measures 10 behaviors (behav.) with every behavior being measured with 2 items. The 7-point # sclae (FSQ) is a uni-dimensional scale and has 18 items. Based on the construct of each questionnaire, we split them evenly into two parts.

	PTS	BREQ-3	PUMP	FSQ
Voice	6 P-items, 5 N-items	6 dim. × 2 items	10 behav. × 1 item	9 items
Web-based	5 P-items, 6 N-items	6 dim. × 2 items	10 behav. × 1 item	9 items

form. As such, we have both open-ended questions (respondents give answers in their words) and close-ended questions (respondents give answers from pre-defined options) with varying rating scales (2-, 5-, and 7-point scales) in two forms (word labels and numerical labels) [35].

To avoid participants answering the same questions twice (one on the web and one via voice), we split each of the four close-ended questionnaires in half. We deployed one half as a conventional web survey and the other half as a voice survey. This approach is inspired by a commonly used method to measure internal consistency - *split-half reliability*. Previous studies on our selected questionnaires have already confirmed that those questionnaires have great split-half reliability, i.e., any halves of those questionnaires are equal, and participants should score similarly in both halves [17, 24, 77, 86]. Further, each of those selected questionnaires has either similar structured or oppositely structured items that allow an even split. For example, two similarly structured items from BREQ-3 are "I exercise because it's fun" and "I enjoy my exercise sessions"; two oppositely structured items from PTS are "(negative) When I think of myself, I think of many shortcomings" and "(positive) I think of myself as a person with many strengths". The allocation of items for each selected questionnaire is shown in Table 1. For the demographic questionnaire, we transform it to be both voice and web-based surveys and use identical questions in both modalities. The split questions can be found in the Appendix A.1 and A.2. In order to be a viable platform to administer voice surveys, responses to voice questionnaires should correlate with responses to web questionnaires.

3.1.2 Survey System Implementation. We deploy the voice survey on Google smart speakers – a widely used speaker platform. We use the Google Actions Builder [1] to implement a Google action called *Monkey Forecast* that administers voice survey surveys described above. The name *Monkey Forecast* was chosen as it can be accurately recognized by Google based on our testing. *Monkey Forecast* was published as an alpha test version, and participants were given the opt-in link. Once they opted in, they automatically have the access to the action. We designed five commands that can directly trigger the five voice questionnaires, i.e., users can invoke any questionnaires by uttering our designed voice commands (see Appendix A.3). Once a voice survey is invoked, the speaker starts by uttering instructions regarding how to answer questions. For example, the instruction for the Binary scale is "For each of the following statements, indicate whether it is yes or no for you. Simply respond with yes or no." Accordingly, the rating scale for 5- and 7-point scales are explained. Except for open-ended questions (DQ), all other questionnaires are designed only to accept the required types of responses (numbers or words).

The voice survey action provides two fallback responses: 1) "no-input" when the speaker does not capture any responses, and 2) "no-match" when the respondent does not provide a valid answer (e.g., the user answers "yes" to a 5-point scale question). Both fallback responses are designed to guide users to "correct" their responses in later attempts. If users trigger "no-input" or "no-match"

[1] https://developers.google.com/assistant/console/builder

errors for the second or the third time, the fallback responses will include example commands that users can use to have the speaker repeat the question ("repeat the last message") or the rating scale ("tell me about the rating scale").

We implement the parallel web survey questionnaires on Qualtrics [2]. We try to keep the format of web surveys as "similar" to voice ones [18]. For example, we use identical instructions for voice and web-based questionnaires. Further, we control the number of labeled points to be the same [19]. We use the fully labeled scales for the Binary and the 5-point W scales, and the polar-point labeled scales for the 5-point # and the 7-point # scales. In particular, as suggested by prior research [43], we design all the rating scales to be only explained once in voice surveys. Accordingly, the scales for web surveys are also presented in the beginning only.

3.1.3 Experimental Design. As aforementioned, we implement five voice surveys and five parallel web surveys with different question types and topics. Thus, we identify two independent variables: 1) modality (voice versus web) and 2) question types (open-ended/binary/5-point #/5-point W/7-point #). We choose to administer different surveys because this can avoid the reuse of questions (i.e., assigning different rating scales to the same questions) and allow participants to experience different question types and provide their preferences on rating scales. On the other hand, we acknowledge that the topics and difficulties of different surveys may be confounding factors. To eliminate the impact of confounding factors as much as possible, we mainly use the notion of rating scales (e.g., 5-point #) rather than survey titles (e.g., BREQ) and remind participants that they should only consider the question type rather than the question content throughout the study. We will discuss this limitation later in Section 6.

We design a within-subject study, where participants are asked to complete both voice survey questionnaires and web survey questionnaires. After they complete each voice (or web) questionnaire, participants need to fill out a short 3-question mini smartphone scale about their experiences with the questionnaire they just completed. Once the mini scale is completed, participants will be instructed to respond to the next voice or web questionnaire. After completing five voice (or web) questionnaires, participants need to complete a 12-question system evaluation form on their smartphones. In the end, we provide participants with twelve in-home scenarios proposed in previous studies [9, 74, 84] (see Table 2), and ask them to choose whether they prefer the voice or the web-browser in each scenario. In total, we collect ten 3-question mini scales, two system evaluation forms, and one modality choice form, which provide us with subjective evaluations of voice and web surveys. We also conduct an exit interview after participants complete all surveys and questionnaires, where we can learn how users experience the new voice survey modality and the traditional web modality.

In addition, we *repeat* the aforementioned study procedure with participants exactly one week after their first session (i.e., each participant experienced the same study session based on their assigned order twice). By repeating the same study procedure twice, we can obtain the *test-retest reliability* of voice surveys and web surveys. We choose to calculate the test-retest reliability as it is commonly used to measure survey reliability in the field of survey methodology [51]. This metric enables us to compare the reliability of voice surveys and conventional web surveys.

The order of five questionnaires may impact participants' performance. The first few voice questionnaires may be more challenging to answer due to unfamiliarity with the speaker modality, whereas participants may feel fatigued in later questionnaires. Also, the difficulty of questionnaires also varies. To mitigate those effects, we designed four questionnaire orders (see Appendix A.4). Participants are randomly assigned one survey order, and they complete web surveys and voice surveys in the same assigned order. Lastly, we further counterbalance the order of survey modalities:

[2] https://www.qualtrics.com

Table 2. Different scenarios for participants to choose survey modality.

You are having breakfast.
You are about to head out.
You are lying in the couch and watching TV.
You are cooking dinner.
You are browsing work-related documents on your personal computer.
You are texting with friends on your phone.
You are doing push-ups.
You are talking to family members.
You are petting your dogs/cats.
You are reading a book.
You are online shopping on your personal computer.
You are doing chores at home.

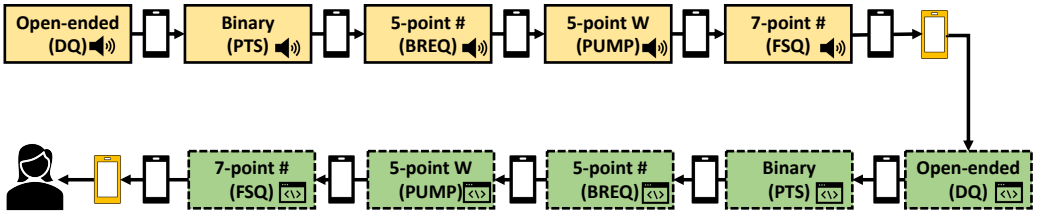


Fig. 1. One example of the experiment procedure. In this example procedure, participants completed voice surveys on smart speakers first (denoted by solid-line squares and speaker icons) and the complete web surveys (denoted by the dash-line squares and browser icons). After each voice (or web) survey, participants complete a 3-question survey experience evaluation on their smartphones (denoted by the black smartphone icon). Once participants complete all five chosen surveys on smart speakers (or on the web), they complete the survey system evaluation form on their smartphones (denoted by the yellow smartphone icon). In the end, researchers conduct an exit interview with participants (denoted by the human icon). To measure the test-retest reliability of surveys, participants are requested to complete the same experiment procedure twice with 1 week apart. To counterbalance the effect of survey orders and modality orders, we designed 8 experiment procedures (4 survey orders × 2 modality orders), which is referenced in the Appendix A.4.

half participants are assigned to finish voice surveys first (voice-web order) while the other half (web-voice order) finish web surveys first. One example study procedure is shown in Figure 1.

Due to COVID-19, we conducted our study over Zoom. We gave participants access to our test Google action and provided them with voice commands that invoke voice surveys. We advertised our study on our university's notice board and LinkedIn, and we also recruited participants through the snowball sampling method. Participants completed our study with their own Google speakers and smartphones and were given a $30 gift card after the completion of our study. We obtained ethics approval from our university's human-subjects committee.

3.2 Participants Demographics

We recruited 24 participants, and all of them are Google smart speaker owners. There are 11 (46%) male and 13 (54%) female participants aged from 19 to 60 years (M = 31.4, SD = 9.0). 4 participants have the highest degree of high school, 9 participants have an undergraduate degree, and 11

participants have a postgraduate degree. All participants are experienced users of smart speakers, and their experiences range from 1 year to 5 years. 21 participants own 1 to 3 smart speakers, while 3 participants own more than 6 smart speakers. Regarding the speaker usage frequency, 20 participants report that they use smart speakers daily or very often, while 4 of them report they rarely use smart speakers. 16 participants used Google Nest Mini, 5 participants used Google Home, 1 participant used the Google Nest Mesh, and 2 participants used the Google Max Hub (we asked them to cover the screen) to complete our study.

3.3 Measures

To investigate whether smart speakers can be a reliable platform to administer voice surveys, we use five measures:

Response Time. We calculate the total completion time for each survey administered by smart speakers and on the web-browser. We also calculate the average response time for answering each question on both modalities. The completion time and the average question response time include not only the user's "thinking and responding time" but also the time for speakers to prompt questions and process responses or the time for web-browsers to respond to user inputs.

Agreement and Correlation. As aforementioned, we split each of the four established questionnaires into two and deployed one half as voice questionnaires and the other half as web questionnaires. The four selected questionnaires all have excellent internal consistency [17, 24, 77, 86], we can check the reliability of voice surveys by comparing the data collected from the two modalities and calculating the agreement between responses. For open-ended questions, three human raters read all participants' responses and checked whether one's voice responses were equal to the corresponding web responses for each question with binary codes: 1 - equal, 0 - not equal. There were occasions where one rater disagreed with the other two raters. For example, to the question - What do you enjoy most about smart speakers?, one participant answered "They make life more convenient for setting timers/alarms, turning lights off etc." on the web and answered "they make life more convenient" to the speaker. Two raters considered these two responses equal, whereas one rater disagreed. In those cases, we adopted the code agreed on by two raters. The inter-rater reliability is 95.3%. Hence, we calculate the percentage of response agreement for each open-ended question. For close-ended questionnaires, we first sum up the total score of each questionnaire and calculate Spearman's rank-order correlation coefficient between the voice questionnaire scores and web questionnaire scores.

Test-retest Reliability. For open-ended questions, similar to the aforementioned, three human raters assessed whether the responses were consistent between two sessions. The code (1 - consistent, 0 - inconsistent) agreed on by two or more raters was chosen, and the inter-rater reliability is 93.7%. Other four surveys' test-retest reliability is measured with the Spearman rank-order correlation and Cohen's kappa. Our hypothesis is that if the voice surveys are reliable, their test-retest reliability results should be comparable to those of web surveys.

Non-differentiation. Satisficing indicates that respondents give "satisfactory" answers rather than "optimal" answers [34]. Common satisficing behaviors include non-differentiation or straight-lining (i.e., give the same answer for a battery of questions), skipping items, abandoning surveys, random selections, rushing to finish surveys, giving minimal words to open-ended questions [4]. Since we collect responses in a semi-controlled study (over Zoom), we do not expect satisficing behaviors, such as skipping items or abandoning surveys, to occur. We adopt the level of non-differentiation of responses of each close-ended questionnaire to infer one satisficing behavior - straight-lining [30, 49]. We use the differential index (ρ) calculation method proposed in [49]. The equation is:

$$\rho = 1 - \sum_{i-1,n} P_i^2 \tag{1}$$

A higher ρ value indicates more differentiation in responses and respondents give more diverse response options. A lower ρ value (close to 0) indicates that respondents almost give identical response options. It should also be noted that ρ is dependent on the length of scales.

Usability. As aforementioned, we ask participants to rate their experiences with each survey questionnaire with a 3-question smartphone scale. The three questions are: 1) I can easily understand those questions, 2) I can easily think of answers for those questions, and 3) Recording my answers is ____(). The first two items are scored using a 5-point Likert scale while the last item requires participants to choose an answer from "Extremely difficult", "Somewhat difficult", "Neither easy nor difficult", "Somewhat easy", and "Extremely easy". The first question measures the *difficulty of understanding*; the second question measures the *difficulty of forming answers*; the last question measures the *difficulty of recording/logging answers*. For each modality, we also ask participants to fill out a 12-question system evaluation form (see Table 8).

4 RESULTS

In this section, we first present quantitative results, including the comparison between voice responses and web responses, test-retest reliability, and the evaluation of response quality. Then, we present the user experience with voice surveys, including results from subjective user evaluation scales and forms and qualitative feedback from exit interviews.

4.1 Response Scale Distributions

Previous studies [19, 73, 90] suggest that voice (e.g., telephone) respondents tend to give more positive answers than do web respondents. Although our participants are both voice respondents and web respondents, we would like to know whether the distribution of their responses to questions on the two modalities would differ. We show the distribution of response rating scales of voice and web surveys in Figure 2. From the figure, it can be observed that the distributions of response scales on the two modalities of our study are comparable. Further, Wilcoxon signed-rank tests also confirmed that the responses from two modalities had no significant differences. Between the two 5-point scales, their distributions are quite different, which may be the result of the survey content.

4.2 Survey Completion Time

We show the survey completion time comparisons in Figure 3. On average, participants took around 60 sec to complete each of the web surveys: DQ = $67'(\pm 20)$, Binary = $55'(\pm 19)$, 5-point # = $59'(\pm 25)$, 5-point W = $56'(\pm 17)$, 7-point # = $51'(\pm 17)$. However, participants took much longer time to complete voice surveys: DQ = $87'(\pm 8)$, Binary = $90'(\pm 9)$, 5-point # = $122'(\pm 21)$, 5-point W = $114'(\pm 21)$, 7-point # = $102'(\pm 13)$. In particular, we observe that for 5-point and 7-point scales, the completion time of voice questionnaires is approximately doubled than that of web questionnaires. On the other hand, we also see a trend in the standard deviation: the completion time of voice questionnaires is less varied than that of web questionnaires except for the 5-point W scale.

Since the four questionnaires are of different lengths, we also calculate the average answering time for each question. The answering time to the first question in each survey on both modalities is removed from the calculation as the voice surveys include instruction on rating scales in the first question. As can be seen from Figure 3, the question-answering time of voice DQ and web DQ is comparable while the question-answering time of other voice questionnaires is significantly longer than that of web questionnaires.

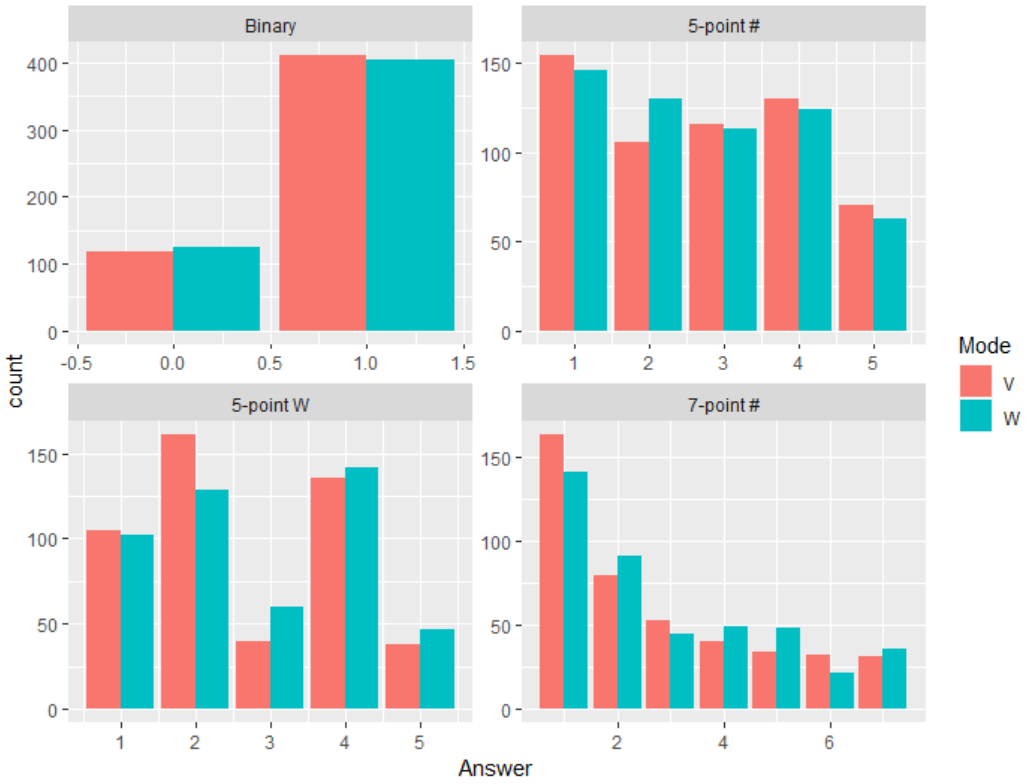


Fig. 2. Distributions of response rating scales of voice surveys (red) and web surveys (blue).

It should be noted that the completion and answering time of voice surveys also include the time cost of interaction errors. For example, participants may not have their responses recognized and recorded the first time. If a no-input error is triggered, the speaker would first prompt the fallback message, and the participant needs to repeat the answer again. So, the variance in answering time of voice questionnaires can be caused by the occurrence of errors and time spent on re-prompts. The relatively low variance in completion time further suggests that most time on voice questionnaires is spent on the speaker prompting and processing. Conversely, as web questionnaires do not trigger any interaction errors, the high variance may actually indicate that some participants were rushing to complete the surveys [4].

4.3 Response Agreement

We calculate the response agreement between voice survey responses and web survey responses. The response agreement of DQ is presented in Table 3. For Q3. Education, one participant gave different expressions of "high school" in voice and web responses, so all raters considered the two responses were not equal. For Q4. Employment, one participant answered "student, working part time" on the web and "student" in the voice. For Q6, languages were sometimes not fully captured by speakers, which resulted in a lower response agreement. In terms of Q7, one participant abbreviated the voice response. Lastly, participants appeared to report differently on Q8. Length of usages. We observed that many participants paused and counted how long they had used smart

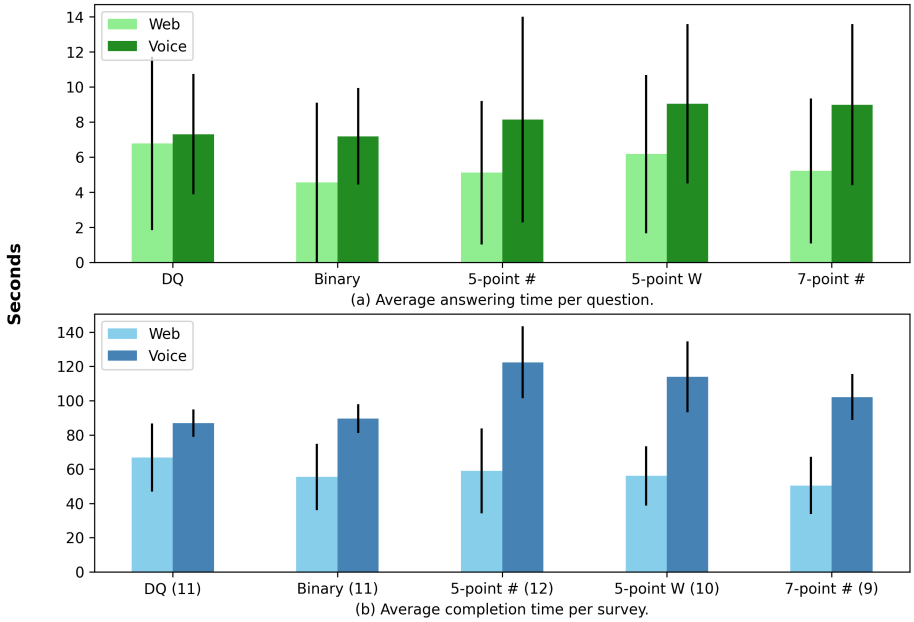


Fig. 3. Survey Completion Time Analysis. The upper figure (a) shows the average time to answer one question for each survey. The lower figure (b) shows the average time to complete one survey.

speakers during the study sessions. So, the disagreements may indicate that it is challenging to answer voice survey questions that require some time to think.

We show Spearman's correlation coefficients ρ of different surveys between modalities in both sessions in Table 4. We also provide the breakdown based on modality orders. Overall, voice responses have a strong, positive correlation with web responses. In particular, we observe that the response agreements of 5-point # (S1: ρ = .77, p <.000, S2: ρ = .89, p < .000) and 7-point # questionnaires are excellent (S1: ρ = .89, p <.000, S2: ρ = .93, p < .000), which indicates that the voice modality has no obvious impact on responses to those questionnaires. However, while the web-voice order group shows significant correlations of all questionnaires, the voice-web order group has non-significant correlations of the Binary questionnaire in S2 (ρ = .50, p = .1) and of the 5-point W questionnaire in S1 (ρ = .33, p = .29). Interestingly, correlations between web responses and voice responses from participants of the voice-web group of all questionnaires except the Binary questionnaire increased in their second session. In particular, those participants noticeably improved their response agreements of 5-point # and 5-point W. These improvements may indicate a learning effect for voice surveys. For participants who were from the web-voice order group, there was no obvious improvement.

4.4 Test-retest Reliability

Another metric to evaluate the voice survey system is the test-retest reliability. In Table 3, we present the test-retest reliability of open-ended questions (DQ). As can be seen from the table,

Table 3. Response agreement (per question item) between web demographic questions and voice demographic questions in two sessions (S1: session 1, S2: session 2), and test-retest reliability of demographic questions on both modalities as indicated by agreement percentage. (The demographic questions can be found in the Appendix)

Demographic	Agreement (%)		Test-retest (%)	
Questions	S1	S2	Voice	Web
Q1. Age	100	100	96	96
Q2. Gender	100	100	100	100
Q3. Education	96	96	100	100
Q4. Employment	96	100	92	96
Q5. Languages	92	96	92	92
Q6. Number of smart speakers	100	100	92	92
Q7. Ways of getting speakers	100	96	96	92
Q8. Length of usages	88	83	67	54
Q9. Frequency of usages	100	100	96	96
Q1. Enjoy most	100	100	96	96
Q11. Biggest complaint	100	100	96	96

Table 4. Correlation coefficient between voice survey questionnaires and web-based survey questionnaires in different orders and in both sessions (S1: session 1, S2: session 2). Bold indicates statistical significance ($p < .05$), and underline indicates no statistical significance.

Order	Voice-Web		Web-Voice		All	
	S1	S2	S1	S2	S1	S2
Binary	**.75**	.50	**.81**	**.82**	**.78**	**.69**
5-point #	**.81**	**.97**	**.77**	**.78**	**.77**	**.89**
5-point W	.33	**.71**	**.79**	**.73**	**.54**	**.77**
7-point #	**.81**	**.92**	**.98**	**.90**	**.89**	**.93**

Q8. Length of usages is the only question that has a low test-retest reliability score. However, participants seem to be more consistent in answering this question in voice than on the web.

For the four close-ended questionnaires, we calculate the correlations of responses between modalities from two sessions. The test-retest reliability results is shown in Table 5. Overall, the test-retest reliability of voice questionnaires and web questionnaires are comparable and both excellent ($\rho > .8$ [38]). We find that the test-retest reliability of the voice binary scale is higher than that of the web binary sale. For the 7-point # questionnaire, participants performed similarly in both modalities. According to Figure 2, the majorities of responses are "1" and "2". The high test-retest reliability could be a result of straight-lining response behaviors, which will be discussed later.

For the 5-point # and 5-point W surveys, however, the modality order appears to impact their test-retest reliability. For voice-web order participants, they responded to web surveys more consistently than voice surveys. Coupled with results from Table 4, we speculate that the lower test-retest reliability of voice forms of 5-point # and 5-point W could still be caused by the learning effect. As participants became more familiar with the voice survey system in the second session, their responses to voice surveys tended to shift closer to their responses to web surveys. On the other hand,

Table 5. Test–retest reliability (Spearman's rank correlation coefficient) and Cohen's kappa for voice and web survey questionnaires. All correlations are statistically significant (p <.05).

	Voice-Web		Web-Voice		All		Cohen's kappa	
Surveys	Voice	Web	Voice	Web	Voice	Web	Voice	Web
Binary	.98	.70	.93	.87	.94	.82	.76	.66
5-point #	.85	.98	.88	.83	.87	.91	.55	.58
5-point W	.77	.90	.87	.75	.83	.85	.41	.45
7-point #	.97	.97	.93	.92	.95	.95	.41	.44

Table 6. Average response lengths of 5 open-ended questions.

	Length		Significance
	Voice	Web	p-value
How did you obtain your smart speakers?	19.3 (10.4)	20.0 (14.8)	.33
How long have you been using smart speakers?	9.1 (4.5)	9.5 (6.43)	.31
How often do you use your smart speakers?	8.4 (3.3)	10.3 (7.5)	.04*
What do you enjoy most about smart speakers?	28.8 (16.5)	37.5 (23.9)	.001***
What is your biggest complaint about smart speakers?	30.7 (16.3)	45.9 (35.4)	.002**

$p^{*}< .05$, $p^{**}< .01$, $p^{***}<.001$.

for web-voice ordered participants, their test-retest reliability of the two 5-point questionnaires is lower on the web than in the voice.

Lastly, to corroborate the test-retest correlations, we also calculated Cohen's kappa for each survey in Table 5. It should be noted that Cohen's kappa is calculated based on the question-level, whereas the correlation is based on the questionnaire-level. Usually, to average out the measurement errors, surveys would include multiple questions/items [69]. Hence, it is understandable that Cohen's kappa tends to be low, and we focus more on the comparisons. We see that the binary survey has the highest test-retest agreement, and in particular, the voice one has a higher kappa than the web one. Comparing the kappa of voice surveys and web surveys, we can see they are quite comparable.

4.5 Evaluation of Response Quality

For open-ended questions, we select the answer length to evaluate the response data quality, and the results can be found in Table 6. While previous studies on smart speakers suggest that VUIs are more advantageous in allowing free-form input [12, 45], we can see that, on average, participants gave shorter answers to voice questions than web questions. In particular, for the last three questions, the web responses are significantly longer than the voice responses.

For close-ended surveys, we use the commonly *Non-differentiation Index* to evaluate the response quality objectively. The results are presented in Table 7. From the calculation equation 1, the index is correlated with the length of rating scales. Therefore, it is understandable that the binary scales have the lowest non-differentiation indices. For the 7-point # questionnaire, however, both voice and web forms of it have lower non-differentiation indices than the two 5-point questionnaires. It could indicate that participants had more satisficing behaviors when answering the 7-point # questionnaire. Nevertheless, through a t-test, we only find that the non-differentiation index of the voice 5-point # questionnaire is significantly higher than that of the web 5-point # questionnaire. We do not observe any significant differences in response differentiation between modalities for other

Table 7. Non-differentiation Index Calculation

	Voice		Web		Significance testing
Surveys	M	SD	M	SD	p-value
Binary	.26	.17	.28	.18	.17
5-point #	.68	.07	.66	.10	.049*
5-point W	.61	.12	.61	.10	.95
7-point #	.49	.24	.46	.29	.40

$p^*< .05$

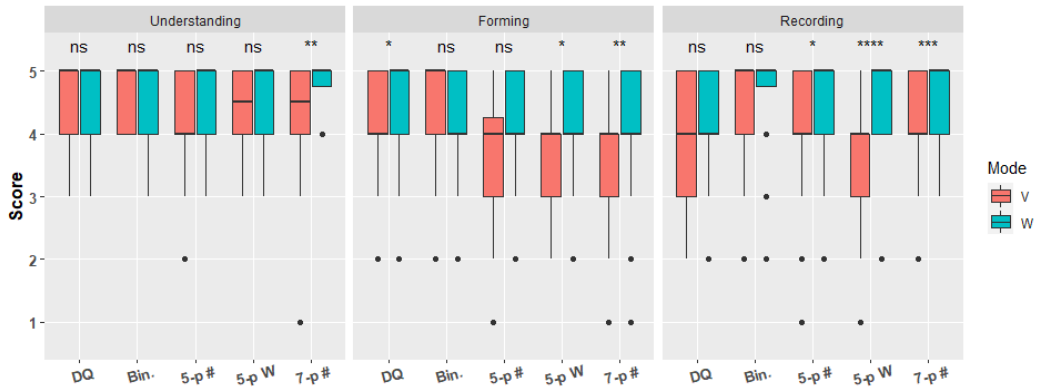


Fig. 4. Subjective evaluations of survey difficulty on three aspects: understanding questions, forming answers, and recording answers.

questionnaires. In other words, the smart speaker modality did not impact participants' response quality to those questionnaires.

4.6 Survey Modality Comparison: Voice versus Web

During the study, participants were asked to fill in the 3-question smartphone scale right after they completed a web or voice questionnaire. This scale aims to measure the subjective difficulty of each survey on three aspects: 1) understanding questions, 2) forming answers, and 3) recording answers. We present the results of the 3-question smartphone scales in Figure 4. A higher score indicates lower difficulty. We use the Wilcoxon signed-rank test to compare the subjective Likert ratings of surveys administered through the voice and the web. In terms of understanding, all voice questionnaires except the 7-point # ($p < .01$) were rated to be equally easy to understand as web surveys. Participants rated similarly for Binary and 5-point # surveys for the forming answer part. As clearly depicted in Figure 4, participants rated that it was more difficult to form responses to DQ ($p < .05$), 5-point W ($p < .05$), and 7-point # ($p < .01$) questionnaires. Lastly, we find that participants rated it was significantly more difficult to record numerical 5-point # ($p < .01$), 7-point # ($p < .001$) answers and 5-point W responses ($p < .000$) via voice than yes/no and open-ended responses on the web. Although we found no significance, as suggested by the boxplot, participants appeared to have varying opinions on the difficulty of recording answers with free-form words.

In our experiment protocol, participants were also asked to fill in a system evaluation form after they completed the set of voice (web) questionnaires. We present the results of system evaluations

in Table 8. Based on the quantitative evaluation results, through the Wilcoxon signed-rank test, we observe that participants scored significantly differently on most questions. First, participants considered the voice survey system was significantly more fun to use (Q1) but was less easy to use (Q2), more difficult to understand (Q4), and slower (Q5) than the web survey system. But participants also acknowledged that they were willing to use the voice survey system in the future. Due to the nature of VUI, we also asked participants to rate how much they would be concerned about the survey confidentiality (Q6). Although the average ratings of both survey systems are low, participants expressed more concerns about the confidentiality of the voice survey system. From the scores of Q7 to Q9, participants rated higher willingness to use the web survey system than the voice survey system if they needed to fill out a survey 1 to 3 times a day. Lastly, although the average ratings are low, participants considered the voice survey system to be more like an ordinary conversation (Q10) and the web survey system to be more like a machine (Q11) and filling in a form (Q12).

Lastly, in our experiment protocol, participants needed to choose the preferred survey modality in different scenarios. We show the breakdowns of user preference in Figure 5. It can be observed that when engaging in physical activities, such as doing push-ups, doing chores, and cooking, participants preferred to use the voice survey system [9, 84]. On the other hand, if engaged in activities that require hearing, talking, and a high cognitive load, participants would prefer the web survey system. In 7 scenarios, voice surveys are preferred to be used. For participants who chose the web survey in those 7 scenarios, we asked about their reasoning. We find that they were mostly concerned about speech recognition and their proximity to smart speakers. For example, one participant explained that she was concerned about the noise produced during cooking that could impact the recognition rate. A few participants also suggested that they would not be close to their speakers if they were heading out or doing chores at home, so it was better to do the web surveys. For scenarios where the majority votes were for the web, we also inquired about participants' reasoning for their choices. Most participants who chose to use the web in those scenarios indicated that they thought the voice surveys would take a longer time and more mental workload to complete while the web surveys could be done quickly. Also, talking to smart speakers can be inconvenient if there is TV sound and rude if they are talking to others. Interestingly, the few participants who preferred the voice in those scenarios indicated that they could easily multi-task with voice surveys. For the two scenarios that already involve the use of computers ("You are browsing work-related documents on your personal computer" and "You are online shopping on your personal computer"), a few participants further suggested that they did not want to switch between web pages when online shopping or working.

4.7 Voice Survey Question Comparisons and Interview Feedback

We present comparisons of question types only in terms of voice surveys in the following.

4.7.1 ***Open-ended versus close-ended.*** From Figure 4, it can be seen that participants considered it more difficult to form responses and rated the recording difficulty of DQ via voice more diversely. Further, many participants also pointed out that they did not think open-ended questions were easy to answer during the interview. For general demographic questions (age, gender, education, and employment status), most participants indicated that those questions were easy and convenient to answer, as there was "only one answer". For questions related to the smart speaker usage history (likes, complaints), some participants said that they needed to "think more" before answering. Some of them tried to keep their answers concise (to improve the speech recognition rate). One participant even pointed out that he imagined a "drop-down menu and its options"

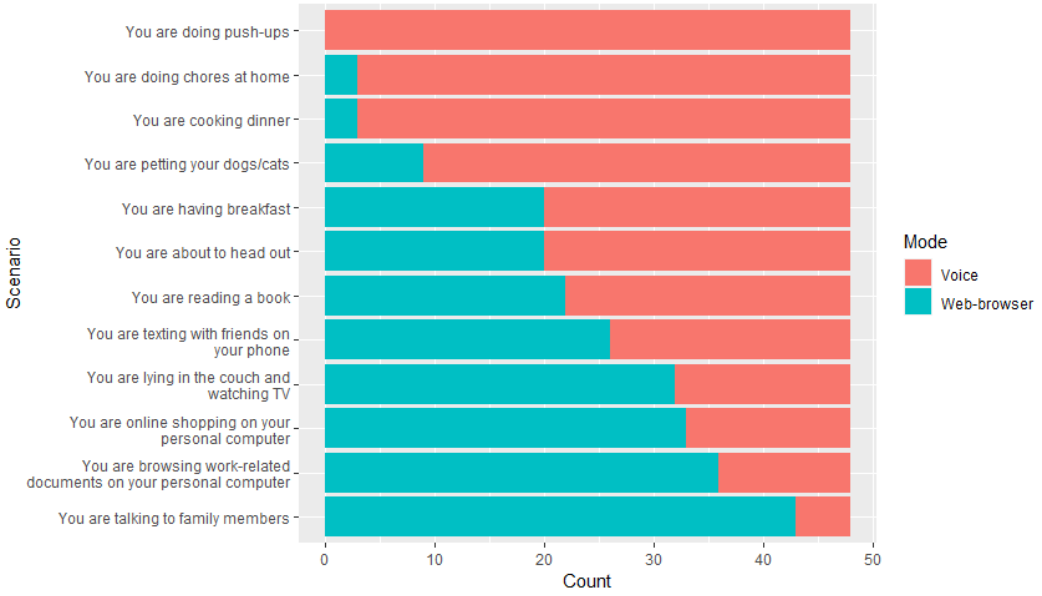


Fig. 5. User preference on voice surveys vs. web surveys in different scenarios.

when responding to some open-ended questions. A few participants also attempted to give longer answers. However, the speaker interrupted their responses while they were still answering.

4.7.2 ***5-point versus 7-point numerical scales.*** When asked about the numerical rating scale, most participants preferred the 5-point scale to the 7-point scale. Many participants explained that they were more familiar with the 5-point Likert scale than the 7-point scale through daily experiences. Hence, most of them had no trouble remembering the 5-point scale and determining what numbers they could choose. However, we noticed that participants still faced two difficulties when answering the 5-point # questionnaire. Firstly, one participant mentioned that she was more used to the scale that started with *strongly agree* while the scale used in our study was in the opposite direction and started with *strongly disagree*. Secondly, the 5-point # survey contained some negatively worded items (e.g., I don't see why I should have to exercise). A few participants reported that it was somewhat difficult to contemplate an answer in the double-negative case. For the 7-point scale, a few participants suggested it was easy to answer as all the questions were basically asking the same thing. However, a few participants were confused about the "middle point" of the scale. Two participants suggested that there was no number that could represent the "neutral" option as they could not answer 3.5 while some other participants thought "5" indicated the "neutral" option. Furthermore, some participants agreed that there was more nuance in the 7-point scale, and they were struggling with the difference between "2" and "3" or "5" and "6". But one participant actually appreciated the minor difference between the two numbers on the 7-point scale.

4.7.3 ***Word versus numerical scales.*** Besides open-ended questions, two other questionnaires also accepted answers with words. The first is the binary scale, to which participants simply answered with yes or no. Many participants suggested that the "yes or no questions are easy as they have only two options" and "it is also quite natural to answer yes (and) no". On the other hand,

Table 8. Quantitative Feedback of Two Survey Systems

Questions	Mean (SD)		W Stat.	p-value
	Voice	Web		
(Q1) It is fun to use the survey system.	4.02 (.73)	3.46 (.90)	371.5	.002**
(Q2) The survey system is easy to use.	3.83 (.88)	4.46 (.58)	31	.000***
(Q3) I'm willing to fill out other surveys on this system in the future.	4.00 (.97)	4.17 (.63)	121.5	.300
(Q4) Understand the survey is _____ (Extremely difficult to Extremely easy)	4.15 (.88)	4.52 (.62)	45	.007**
(Q5) Completing the survey is _____ (Extremely slow to Extremely fast)	3.46 (1.17)	4.17 (.78)	104.5	.001**
(Q6) If you are answering the survey at home, will you be concerned that your response may not be kept 100% confidential?	2.58 (.94)	2.18 (.89)	186	.001**
(Q7) If you are asked to fill out a survey 1 to 3 times a day, will you be willing to use this system?	3.17 (1.23)	3.52 (1.15)	84	.050*
(Q8) If you are asked to fill out a survey 3 to 5 times a day, will you be willing to use this system?	2.85 (1.27)	3.19 (1.16)	184	.092
(Q9) If you are asked to fill out a survey 5 to 10 times a day, will you be willing to use this system?	2.25 (1.31)	2.43 (1.18)	241	.467
(Q10) How much is this survey system like an ordinary conversation?	2.69 (1.13)	1.81 (1.00)	507	.000***
(Q11) How much is this survey system like dealing with a machine?	3.19 (.98)	3.54 (.99)	190	.010**
(Q12) How much is this survey system like filling in an form?	2.67 (1.14)	4.44 (.71)	0	.000***

$p^{*} < .05$, $p^{**} < .01$, $p^{***} < .001$.

some participants considered that some of the questions were quite deep and the dichotomous choice was somehow too extreme. Nevertheless, the binary scale was still considered to be the easiest by most participants. Regarding the 5-point W survey, as can be seen in Table 4, Table 5, and Figure 4, participants not only performed relatively poorly on this survey but rated it to be difficult to record answers. During the exit interview, the 5-point W scale was also not preferred by most participants. Some participants mentioned that "it is difficult to remember the exact words", and others further suggested that "it is literally more difficult to say the word than the number". In fact, more than one participants answered "somewhat disagree" and "somewhat agree" during the study, which also indicated that remembering the 5-point W scale was challenging. Lastly, a few participants also suggested that answering "strongly disagree" or "strongly agree" was in fact not natural as "people usually would not say *I strongly agree* during daily conversations". Conversely, if to give extreme answers, participants expressed they had no difficulty in answering "1" or "5" compared to "strongly disagree" or "strongly agree".

Table 9. Interaction error rate of each voice survey.

DQ	Binary	5-point #	5-point W	7-point #	**Total**
3.79%	2.65%	4.68%	9.79%	7.64%	**5.54%**

4.8 Voice Survey Interactions and Errors

Lastly, we present the usability measures of smart speakers as the voice survey administration platform. We find that some voice surveys are pre-maturely terminated due to connection errors. A few participants reported that they "lost" the speaker during the study session. For example, the speaker remained silent to user responses and failed to respond back or prompt an error message "hmm, something went wrong". Another interesting case is that one participant tried to turn up the volume of the speaker, which somehow caused the speaker to stop the survey. In these cases, we would instruct participants to re-initiate the voice surveys.

As mentioned in the Methods section, we construct the error prompts to be informative and contain "repair commands" ("repeat the message" and "tell me about the rating scale"). If participants encounter the no-input or no-match error more than 2 times, they can learn repair commands. In total, the Repeat command was used 3 times by 3 participants, and the Rating Scale command was used 6 times by 5 participants (5-point #: 3, 5-point W: 2, 7-point #: 1). On the other hand, 2 participants, who did not trigger error prompts that include repair commands, chose to re-initiate the surveys as they forgot the rating scale and tried to resolve this issue by re-invoking the survey. In 14 out of 48 sessions from 24 participants, 12 participants had to re-initiate one (or two) voice surveys. We only used the last completed survey responses for the data analysis below.

Finally, we provide a summary of errors that occurred during the study. We calculate the interaction error rate for each voice survey in Table 9. Participants encountered most interaction errors when responding to the 5-point W, which, interestingly, is consistent with the highest user-rated difficulty of this scale. Also, we notice there still exist prevalent transcription errors in our dataset. For open-ended questions, the word error rate (WER) of transcribed responses is 15.2%. Therefore, we manually transcribed participants' responses again and modified all incorrect transcriptions for our analysis. The most common error is that the speaker cannot correctly recognize the word "male". For all male participants, only one participant's "male" was recognized once, all others' "male" was recorded as "mail", "miel", or "Mel". Transcription errors also occurred frequently in the language question. For participants who know more than three languages, their speakers sometimes failed to capture all the languages. Similarly, when recording longer responses, speakers also tended only to capture part of participants' answers (e.g., "sometimes it doesn't understand and I have to repeat" was recorded as "sometimes I understand"). In terms of close-ended questions, since the answer is limited, we mainly observe one type of error: doubled- or tripled numerical answers caused by participants repeating their answers (e.g., answered "1" for two times, then "11" was recorded).

5 DISCUSSION

5.1 Implications for voice survey formats

Our results suggest that the response agreement and the test-retest reliability of voice and web surveys are overall comparable. Also, the voice modality even improved the response differentiation rate of the 5-point Likert scale. Therefore, we think voice survey questionnaires administered by smart speakers can be an effective alternative to traditional web survey questionnaires. However, we do find that some question formats may be more suitable to be deployed as voice surveys on smart speakers than others.

First, the open-ended questions are not as suitable as we expected. The voice survey is good for general demographic questions, such as age, gender, and education, which is aligned with findings from previous studies on IVR surveys [29, 63]. For the rest questions, it is not the case. While speaking is faster and more effortless than typing [64], the lengths of responses collected via voice were actually shorter than those collected on the web. In a few rare cases, some participants were interrupted by the speaker during the experiment and could not record their full answers (this does not affect the prior conclusion). In terms of the answer contents, many participants did not respond freely, and some constructed their answers carefully as they were all well aware of the limitation of speech recognition. Lastly, we also notice that transcriptions of free-form responses contain many recognition errors that need to be corrected manually [45]. Further, we show that the WER is 15.2% for these free-form responses. Modifying incorrect transcriptions does not only require manual efforts but may also require audio recordings. Therefore, we do not think voice surveys should include too many open-ended questions that may trigger long responses [84].

Second, consistent with prior literature [35], we do find that the length of rating scales impacts the reliability and quality. The binary voice questionnaire appeared to have a good consistency with its web counterpart and even better test-retest reliability and kappa [63]. The recognition of yes/no is also more robust to errors. Although some participants did complain that this scale was too absolute with no middle ground, we still think this format can work well in future voice questions. For longer-length rating scales, prior literature suggests that the 7-point scale is the optimal one for paper- and web-based surveys for its higher validity and reliability [35, 36, 55]. In this study, although the 7-point # appears to have great response agreement and test-retest reliability, which could be a result of satisficing. As can be seen in Table 7, the 7-point # has a lower non-differentiation index than the 5-point ones, which indicates less variance in responses. Since this questionnaire is uni-dimensional and measures people's fear towards spiders, for participants who are not arachnophobic, they would give straight-lining answers (i.e., answer "1" for all questions) in both modalities. This explains the higher agreement between modalities and test-retest reliability. Compared with the 5-point scale, the 7-point scale is not familiar to most participants and has more nuance and smaller differences between adjacent options. When lacking visual references, the 7-point scale is more cognitively difficult for respondents to use [36]. For the two 5-point scales, the word one performed worse and was considered more troublesome to use. Further, according to Table 9, 5-point W is the most error-prone one. Therefore, we suggest that the 5-point # scale may be the most appropriate one for working in the voice modality.

5.2 Implications for voice survey designs

Although we focused on investigating the question types for voice survey questionnaires, we also received some user feedback on the wording of the questions. In many surveys and questionnaires, negatively worded items are included as they may reduce participants' acquiescing, satisficing, and inattention behaviors [2]. For example, the commonly used System Usability Scale [6] in the CSCW community contains half negatively worded items. However, our participants report that those negative questions were difficult to answer in voice. In particular, some participants struggled with the direction of the rating scale, which may be due to the lack of visual presentation [19, 63, 70].

In Figure 3, we show the question-answering time and the completion time of each web and voice survey. Obviously, voice surveys take a longer time to complete than web surveys. In particular, we observe doubled completion time for voice surveys compared to web surveys. To this end, previous studies suggest that the time cost of participation and the survey length can impact the response rate [35]. Therefore, researchers aiming to use voice surveys or mix-modality surveys should consider the factor of completion time. Additionally, when transforming paper or web-based surveys to be voice ones, researchers can consider reducing the length of surveys to half.

Lastly, we note that there appears to be a learning effect for the voice survey system. Although voice-based surveys have been widely implemented through IVR systems, especially the use of customer service call centers [66], most of our participants seem to consider voice surveys on smart speakers to be a new "thing". As we have the test-retest study design, we observe that many participants became more used to the speaker survey system. In Table 4, participants in the voice-web order have a lower response agreement of 5-point # and 5-point W surveys than those in the web-voice order in the first session. Web-voice order participants may be already familiar with the surveys and various rating scales; on the other hand, participants in the voice-web order were exposed to new surveys in a new modality. The improvement of response agreement in the second session for the voice-web order participants, particularly for the two 5-point surveys, indicates the learning effect of the new survey modality. During the interview, many participants also acknowledged the notion of "familiarity". For example, when asked "do you think you'll perform better next time?", one participant suggested that "Yes, I think so. Yeah, because I was not used to (the system), but once I get used to the system, I know what's expecting." In the second session's interview, many people also stated that they knew what they needed to do this time. Therefore, in line with other works [43, 45], we also suggest researchers consider adding training questions in their voice survey design to help respondents get used to the smart speaker survey system as well as the used rating scales.

5.3 Smart speakers as a new survey administration modality

Due to the increasing popularity of commercial smart speakers, we envision that smart speakers can be a good platform for survey administration to collect user data in home [42, 44, 45]. Previous studies on IVR surveys found that respondents tended to emulate the speaking styles of the voice, it is therefore suggested that the voice prompts could be recorded slowly but clearly and preferably by native speakers [40]. Recent technologies have made generated speech more human-like [1]. In our study, the voice surveys were prompted by participants' personal Google speakers. We noticed that different users had different settings for their speakers. For example, some participants configure their speakers to have different accents, and a few participants had the accessibility feature enabled. Hence, unlike IVR surveys that are identically and centrally distributed [19], smart speakers, by default, can prompt surveys with personalized voices in different genders, accents, and even speech rates.

Considering that smart speakers are usually located in people's homes, we proposed 12 common in-home scenarios and investigated people's preference for survey modality in those scenarios. As can be seen in Figure 5, combining all the votes, the voice was actually chosen more than the web. On the other hand, participants also rated voice questionnaires to be more difficult to understand and answer in Figure 4 and Table 8. Such discrepancy may be due to some limitations in smart speakers' functionality. Here, we summarize the interaction obstacles and difficulties with smart speakers.

We consider our study was conducted in a semi-controlled environment, where participants stayed in their homes without any obvious noises, used their own smart speakers, and stayed in close proximity to those speakers. However, there are still two interaction obstacles that negatively impact user experience. The first one is the pre-mature termination of some voice surveys. As indicated in subsection 4.8, about 10% surveys were ended unexpectedly. In those cases, the smart speakers somehow appeared disconnected and could not proceed with our survey action. We had to request participants to invoke the survey again. While it is unknown which part of the cloud pipeline was causing the speaker to be temporarily out-of-service, it was not uncommon. We find this "disconnection" issue seems to be a major factor that impacts the smoothness and user experience of voice surveys. The other obstacle is caused by the speech recognition limitation. For

both open-ended and close-ended questions, smart speakers struggled to misrecognize common words (e.g., "male" was mostly recognized as "mail") and triggered more than 5% interaction errors in total. A few participants had to answer more than twice to get their responses recognized for some questions. There were also occasions where the Google speaker responded to some phrases incorrectly. One instance is that the speaker failed to recognize "28" from "I'm 28" and another instance is that the speaker failed to record "every second day" and kept re-triggering the survey action. These ASR failures are not common yet confusing and have caused participants to re-initiate voice surveys. When being asked about whether interaction errors are bothering them, many participants commented that one time (error) would be okay, but more than two times would be annoying. We would assume that more unstable environmental factors (e.g., noise, other people's presence) exist if people are asked to respond to voice surveys in day-to-day life [84]. Therefore, before mass-deploying voice surveys, researchers should consider measuring users' home network conditions [60], minimizing ASR errors by designing better surveys [29], and evaluating voice questionnaire designs with people with different accents [33, 54].

Besides the limitations in the speech recognition of smart speakers, participants also expressed concerns about the conversational aspect of voice surveys. As current smart speakers only serve as a front-end while all the other processes are completed in the cloud [58], the interaction is not always smooth. Many participants complained about the time gaps between their answers and the speaker's next question. For example, one mentioned that "some gaps are longer, which made me wonder if it was still working". Another participant reported that "there is no sign whether my responses are recorded or not". Interestingly, one participant was actually surprised that he could wait longer before answering. It appears that most participants have no idea of how long it takes for the speaker to respond. When facing longer gaps, participants would either try to check the lighting to infer whether the speaker is still working or simply repeat their answers. We speculate that although all of our participants are experienced smart speaker users, they still lack the experience of engaging in multi-turn conversations with their speakers [5]. This *uncertainty* may stop some people from using the voice survey system.

The clumsiness of voice surveys is also caused by the architecture of the Google platform. In this study, we used Google Actions Builder to build our survey action. With the default flow, questionnaires in our action can only be invoked and started from the first question. In other words, users cannot leave a voice survey halfway and come back to it later. Some participants pointed out this shortcoming and said that they would not be willing to fill out a very long voice questionnaire if they had to complete it all at once. We also noticed that some participants tried to be efficient and answer once the speaker stopped talking, which, somehow, triggered no-input errors. We speculate that this is because the answer was short, and the user's speech had a small, imperceptible overlap with the speaker's speech. Currently, the smart speaker cannot receive user input when it is speaking [89]. While such design may have its own advantages, it is time-consuming when users face interaction errors. In the case of no-input errors, the moment the speaker prompts the fallback message, users will know the error and can attempt to repeat their responses. However, the current design forces users to wait until the speaker stops prompting. This ultimately increases the time cost of voice surveys, and many participants found this annoying. Lastly, the smart speaker is still "not human enough" [83]. A participant reported that the computer-generated speech was more difficult to follow than a real human's speech. Nevertheless, we believe the usability of the voice survey system can be significantly improved if smart speakers can enable more flexible communication repairs and more natural conversations.

6 LIMITATIONS AND FUTURE WORK

While there are many smart speaker owners worldwide, we find it difficult to recruit participants who are willing to participate in a 2-hour study, especially during COVID-19. Therefore, our sample size is limited. To collect as much data as possible, we designed a within-subjects study where each participant experienced both survey modalities and responded to five different surveys with varying answer options. One recent study that compared the voice survey with the pen&paper survey also adopted a within-subjects design [45]. They randomized the question order in the two modalities to mitigate the possibility of participants' familiarity with the questionnaire. However, this approach could be unreliable as participants may still remember some questions and respond without "thinking". To avoid that, we split four established questionnaires into halves and implemented those halves on the speaker and the web. We acknowledge that there are confounding factors in our study design. First, although all established questionnaires have excellent split-half reliability [77], the voice questions and web questions are still not entirely identical. So, respondents are not supposed to answer identically across modalities. Second, those surveys are of different topics and difficulties, which could also become a confounding factor. To eliminate the impact of question contents, we emphasized to participants that they should only consider the format of questions rather than the content when they answer the 3-question scales and during their exit interviews. Nevertheless, it is undeniable that the reliability and response agreements are compromised to some extent.

As aforementioned, we noticed a learning effect of voice surveys, particularly for participants in the voice-web order group. The learning effect may negatively impact the effectiveness of test-retest reliability measurement. As participants became more familiar with the rating scale and gave more thought to the questions, they may give different answers to the same questions, which lowered the test-retest consistency. Lastly, there was an observer effect during the survey administration process. A few participants actually pointed out that they were slightly uncomfortable answering some personal (e.g., age) and attitude (e.g., "I frequently compare myself to others.") questions in front of our researchers. With the presence of an observer, participants may opt to choose more "positive" answers without thinking more deeply, or they may attempt to give longer answers to open-ended questions [22, 57]. The response data quality may be compromised when people are asked to respond in the wild.

Although this study design has limitations, we find our design allows participants to experience all types of questions and provide their opinions and preferences to different question formats, which is essential for us to learn about user perceptions at this exploratory stage. Alternative between-subjects designs, where different respondents answer one identical survey with different formats of questions, may yield a more rigorous measurement of the reliability and suitability of rating scales. However, in that design, a larger sample size will be required and participants cannot directly compare different rating scales. Hence, we think both within- and between-subjects design studies are crucial to deepen our understanding of voice surveys on smart speakers. Also, future research should integrate training for participants to mitigate the learning effect. Lastly, to evaluate how smart speakers perform as a new survey administration platform in the wild, it is necessary to measure the response rate, drop-out rate, and data quality of voice surveys [19, 29]. We designed our Google action and published it as a test version for our participants. To mass distribute smart speaker surveys, people need to be aware of this new type of survey. Researchers can look into how to "send" voice surveys to speaker owners. Maybe voice survey VUI applications (e.g., Google actions and Alexa skills) can be published publicly and send to users through daily notifications [3].

[3]https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/notifications-overview.html

Lastly, researchers can examine ASR errors in voice surveys, as we believe the willingness to adopt new survey modalities depends largely on the interaction experience [29].

7 CONCLUSION

In this study, we investigated what types of questions are suitable to be deployed in voice surveys administered by smart speakers. With a within-subject design, we compared the response agreement between voice and web responses and the test-retest reliability of five different surveys. Our results suggest that overall voice surveys administered by smart speakers are comparable to web surveys in terms of response validity and reliability. Among the five tested surveys, we find participants performed the best with the binary scale and slightly worse with the 5-point W scale, and they also rated the 5-point W scale more difficult to use. The numerical scales, on the other hand, were considered easier to choose from. While we originally expected that open-ended voice questions might incite longer and higher-quality responses, the data collected via voice was actually not as long as that collected on the web. Participants also considered recording free-form answers to be troublesome as they needed to speak "recognizable" words. Based on empirical data and qualitative user feedback, we recommend researchers use the binary scale and 5-point numerical scale for voice surveys and integrate some training for respondents in the future. Lastly, we reflect on the limitations of our study and suggest that future work should recruit more respondents and evaluate voice surveys with both between- and within-subjects designs.

ACKNOWLEDGMENTS

This work is partially funded by ARC Discovery Project DP190102627, NHMRC grants 1170937 and 2004316.

REFERENCES

[1] [n.d.]. WaveNet: A generative model for raw audio. https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

[2] Mike Allen. 2017. *The SAGE encyclopedia of communication research methods.* Sage Publications.

[3] Duane F Alwin. 1992. Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological methodology* (1992), 83–118.

[4] Scott Barge and Hunter Gehlbach. 2012. Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education* 53, 2 (2012), 182–200.

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] S Tamer Cavusgil and Lisa A Elvey-Kirk. 1998. Mail survey response behavior: A conceptualization of motivating factors and an empirical study. *European journal of marketing* (1998).

[8] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies* 139 (2020), 102410.

[9] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello there! is now a good time to talk? Opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–28.

[10] Dipanjan Chakraborty, Indrani Medhi, Edward Cutrell, and William Thies. 2013. Man versus machine: evaluating IVR versus a live operator for phone surveys in India. In *Proceedings of the 3rd ACM Symposium on Computing for Development.* Association for Computing Machinery, New York, NY, USA, 1–9.

[11] Ti-Chung Cheng, Tiffany Wenting Li, Yi-Hung Chou, Karrie Karahalios, and Hari Sundaram. 2021. " I can show what I really like." Eliciting Preferences via Quadratic Voting. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–43.

[12] Jane Chung, Michael Bleich, David C Wheeler, Jodi M Winship, Brooke McDowell, David Baker, and Pamela Parsons. 2021. Attitudes and Perceptions Toward Voice-Operated Smart Speakers Among Low-Income Senior Housing Residents: Comparison of Pre-and Post-Installation Surveys. *Gerontology and Geriatric Medicine* 7 (2021), 23337214211005869.

[13] Richard L Clayton and Debbie LS Winter. 1992. Speech data entry: results of a test of voice recognition for survey data collection. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-* 8 (1992), 377–377.

[14] John Dawes. 2008. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International journal of market research* 50, 1 (2008), 61–104.

[15] Don C Des Jarlais, Denise Paone, Judith Milliken, Charles F Turner, Heather Miller, James Gribble, Qiuhu Shi, Holly Hagan, and Samuel R Friedman. 1999. Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: a quasi-randomised trial. *The Lancet* 353, 9165 (1999), 1657–1661.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Ed Diener, Derrick Wirtz, Robert Biswas-Diener, William Tov, Chu Kim-Prieto, Dong-won Choi, and Shigehiro Oishi. 2009. New measures of well-being. In *Assessing well-being*. Springer, 247–266.

[18] Don A Dillman and Leah Melani Christian. 2005. Survey mode as a source of instability in responses across surveys. *Field methods* 17, 1 (2005), 30–52.

[19] Don A Dillman, Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L Messer. 2009. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social science research* 38, 1 (2009), 1–18.

[20] Tilman Dingler, Dominika Kwasnicka, Jing Wei, Enying Gong, and Brian Oldenburg. 2021. The Use and Promise of Conversational Agents in Digital Health. *Yearbook of Medical Informatics* 30, 01 (2021), 191–199.

[21] Radhika Garg and Subhasree Sengupta. 2020. He is just like me: a study of the long-term use of smart speakers by parents and children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–24.

[22] Khalil G Ghanem, Heidi E Hutton, Jonathan M Zenilman, Rebecca Zimba, and Emily J Erbelding. 2005. Audio computer assisted self interview and face to face interview modes in assessing response bias among STD clinic patients. *Sexually transmitted infections* 81, 5 (2005), 421–425.

[23] Moshe M Givon and Zur Shapira. 1984. Response to rating scales: A theoretical model and its application to the number of categories problem. *Journal of Marketing Research* 21, 4 (1984), 410–419.

[24] Katharina Graben, Bettina K Doering, Franziska Jeromin, and Antonia Barke. 2020. Problematic mobile phone use: Validity and reliability of the Problematic Use of Mobile Phone (PUMP) Scale in a German sample. *Addictive behaviors reports* 12 (2020), 100297.

[25] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I Am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376320

[26] Allyson L Holbrook, Melanie C Green, and Jon A Krosnick. 2003. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly* 67, 1 (2003), 79–125.

[27] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[28] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 143–152.

[29] Aman Khullar, Priyadarshi Hitesh, Shoaib Rahman, Deepak Kumar, Rachit Pandey, Praveen Kumar, Rajeshwari Tripathi, Prince Prince, Ankit Akash Jha, Himanshu Himanshu, et al. 2021. Costs and Benefits of Conducting Voice-based Surveys Versus Keypress-based Surveys on Interactive Voice Response Systems. In *ACM SIGCAS Conference on Computing and Sustainable Societies*. Association for Computing Machinery, New York, NY, USA, 288–298.

[30] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[31] Bret Kinsella. 2019. Loup Ventures says 75% of U.S. households will have smart speakers by 2025, Google to surpass Amazon in market share. https://voicebot.ai/2019/06/18/loup-ventures-says-75-of-u-s-households-will-have-smart-speakers-by-2025-google-to-surpass-amazon-in-market-share/

[32] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference*. Association for Computing Machinery, New York, NY, USA, 881–894.

[33] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[34] Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5, 3 (1991), 213–236.
[35] Jon A Krosnick. 2018. Questionnaire design. In *The Palgrave handbook of survey research*. Springer, 439–455.
[36] Jon A Krosnick and Matthew K Berent. 1993. Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science* (1993), 941–964.
[37] Jon A Krosnick, Sowmya Narayan, and Wendy R Smith. 1996. Satisficing in surveys: Initial evidence. *New directions for evaluation* 1996, 70 (1996), 29–44.
[38] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[39] Kwan Min Lee and Jennifer Lai. 2005. Speech versus touch: A comparative study of the use of speech and DTMF keypad for navigation. *International Journal of Human-Computer Interaction* 19, 3 (2005), 343–360.
[40] Adam Lerer, Molly Ward, and Saman Amarasinghe. 2010. Evaluation of IVR data collection UIs for untrained rural users. In *Proceedings of the first ACM symposium on computing for development*. Association for Computing Machinery, New York, NY, USA, 1–8.
[41] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).
[42] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: Shaping Consistent Exercise Experience by Complementing a Mobile App with a Smart Speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376616
[43] Kelly L'Engle, Eunice Sefa, Edward Akolgo Adimazoya, Emmanuel Yartey, Rachel Lenzi, Cindy Tarpo, Nii Lante Heward-Mills, Katherine Lew, and Yvonne Ampeh. 2018. Survey research with a random digit dial national mobile phone sample in Ghana: methods and sample quality. *PloS one* 13, 1 (2018), e0190902.
[44] Raju Maharjan, Per Bækgaard, and Jakob E Bardram. 2019. " Hear me out" smart speaker based conversational agent to monitor symptoms in mental health. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 929–933.
[45] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *CUI 2021-3rd Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–11.
[46] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2014. Capturing the mood: facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. Association for Computing Machinery, New York, NY, USA, 1082–1094.
[47] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Association for Computing Machinery, New York, NY, USA, 903–916.
[48] David Markland and Vannessa Tobin. 2004. A modification to the behavioural regulation in exercise questionnaire to include an assessment of amotivation. *Journal of Sport and Exercise Psychology* 26, 2 (2004), 191–196.
[49] John A McCarty and Larry J Shrum. 2000. The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly* 64, 3 (2000), 271–298.
[50] Lisa J Merlo, Amanda M Stone, and Alex Bibbey. 2013. Measuring problematic mobile phone use: development and preliminary psychometric properties of the PUMP scale. *Journal of addiction* 2013 (2013).
[51] Elizabeth T Miller, Dan J Neal, Lisa J Roberts, John S Boer, Sally O Cresskr, Jane Metrik, and G Alan Marlatt. 2009. Test-retest reliability of alcohol measures: is there a difference between internet-based assessment and traditional methods? (2009).
[52] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.
[53] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. Number 47. University of Illinois press.
[54] Debajyoti Pal, Chonlameth Arpnikanondt, Suree Funilkul, and Vijayakumar Varadarajan. 2019. User experience with smart voice assistants: the accent perspective. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.
[55] Josh Pasek and Jon A Krosnick. 2010. Optimizing survey questionnaire design in political science. In *The Oxford handbook of American elections and political behavior*. Oxford University Press.
[56] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural India. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 51–54.

[57] Theresa E Perlis, Don C Des Jarlais, Samuel R Friedman, Kamyar Arasteh, and Charles F Turner. 2004. Audio-computerized self-interviewing versus face-to-face interviewing for research data collection at drug abuse treatment programs. *Addiction* 99, 7 (2004), 885–896.

[58] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[59] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. " Phantom Friend" or" Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[60] Aung Pyae and Tapani N Joelsson. 2018. Investigating the usability and user experiences of voice user interface: a case of Google home smart speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Association for Computing Machinery, New York, NY, USA, 127–131.

[61] Ling Qiu, Bethany Kanski, Shawna Doerksen, Renate Winkels, Kathryn H Schmitz, and Saeed Abdullah. 2021. Nurse AMIE: Using Smart Speakers to Provide Supportive Care Intervention for Women with Metastatic Breast Cancer. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.

[62] Juan C Quiroz, Tristan Bongolan, and Kiran Ijaz. 2020. Alexa depression and anxiety self-tests: a preliminary analysis of user experience and trust. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 494–496.

[63] Shan M Randhawa, Tallal Ahmad, Jay Chen, and Agha Ali Raza. 2021. Karamad: A Voice-based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[64] Melanie Revilla, Mick P Couper, Oriol J Bosch, and Marc Asensio. 2020. Testing the use of voice input in a smartphone web survey. *Social Science Computer Review* 38, 2 (2020), 207–224.

[65] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior* 126 (2022), 107034.

[66] George Robinson and Clive Morley. 2006. Call centre management: responsibilities and performance. *International Journal of Service Industry Management* (2006).

[67] John P Robinson, Phillip R Shaver, and Lawrence S Wrightsman. 1999. *Measures of political attitudes*. Academic Press.

[68] Steven J Rosenstone, John Mark Hansen, and Donald R Kinder. 1986. Measuring change in personal economic well-being. *Public Opinion Quarterly* 50, 2 (1986), 176–192.

[69] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3371382.3380739

[70] Norbert Schwarz, Fritz Strack, Hans-J Hippler, and George Bishop. 1991. The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology* 5, 3 (1991), 193–212.

[71] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. Healthline: Speech-based access to health information by low-literate users. In *2007 International Conference on Information and Communication Technologies and Development*. IEEE, 1–9.

[72] Eunjung Shin, Timothy P Johnson, and Kumar Rao. 2012. Survey mode effects on data quality: Comparison of web and mail modes in a US national panel survey. *Social Science Computer Review* 30, 2 (2012), 212–228.

[73] Alicia D Simmons and Lawrence D Bobo. 2015. Can non-full-probability internet surveys yield useful data? A comparison with full-probability face-to-face surveys in the domain of race and social inequality attitudes. *Sociological Methodology* 45, 1 (2015), 357–387.

[74] Ulla Sonn, Kristina Törnquist, and Elisabeth Svensson. 1999. The ADL taxonomy-from individual categorical data to ordinal categorical data. *Scandinavian Journal of Occupational Therapy* 6, 1 (jan 1999), 11–20. https://doi.org/10.1080/110381299443807

[75] Venkat Srinivasan and Amiya K Basu. 1989. The metric quality of ordered categorical data. *Marketing Science* 8, 3 (1989), 205–230.

[76] Ayushi Srivastava, Shivani Kapania, Anupriya Tuli, and Pushpendra Singh. 2021. Actionable UI Design Guidelines for Smartphone Applications Inclusive of Low-Literate Users. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.

[77] Jeff Szymanski and William O'Donohue. 1995. Fear of spiders questionnaire. *Journal of behavior therapy and experimental psychiatry* 26, 1 (1995), 31–34.
[78] Roger Tourangeau and Kenneth A Rasinski. 1988. Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin* 103, 3 (1988), 299.
[79] Priyamvada Tripathi and Winslow Burleson. 2012. Predicting creativity in the wild: Experience sample and sociometric modeling of teams. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. Association for Computing Machinery, New York, NY, USA, 1203–1212.
[80] Charles F Turner, Leighton Ku, Susan M Rogers, Laura D Lindberg, Joseph H Pleck, and Freya L Sonenstein. 1998. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* 280, 5365 (1998), 867–873.
[81] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.
[82] Morgan Vigil-Hayes, Ann Futterman Collier, Shelby Hagemann, Giovanni Castillo, Keller Mikkelson, Joshua Dingman, Andrew Muñoz, Jade Luther, and Alexandra McLaughlin. 2021. Integrating cultural relevance into a behavioral mHealth intervention for Native American youth. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–29.
[83] Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.
[84] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding User Perceptions of Proactive Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.
[85] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15.
[86] Philip M Wilson, Wendy M Rodgers, and Shawn N Fraser. 2002. Examining the psychometric properties of the behavioral regulation in exercise questionnaire. *Measurement in physical education and exercise science* 6, 1 (2002), 1–21.
[87] Philip M Wilson, Wendy M Rodgers, Christina C Loitz, and Giulia Scime. 2006. "It's Who I Am… Really!"The importance of integrated regulation in exercise contexts 1. *Journal of Applied Biobehavioral Research* 11, 2 (2006), 79–104.
[88] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.
[89] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.
[90] Cong Ye, Jenna Fulton, and Roger Tourangeau. 2011. More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly* 75, 2 (2011), 349–365.

A APPENDIX

A.1 Questionnaires administered on the web

1. Demographic questions

(1) What is your age? (Age)
(2) What is your gender? (Gender)
(3) What is the highest degree or level of school you have completed? (Education)
(4) What is your current employment status? (Employment)
(5) Which languages are you capable of speaking fluently? (Language)
(6) How many smart speakers do you own? (Number of smart speakers)
(7) How did you obtain your smart speakers? (Ways of getting speakers)
(8) How long have you been using smart speakers? (Length of usages)
(9) How often do you use your smart speakers? (Frequency of usages)
(10) What do you enjoy most about smart speakers? (Enjoy most)
(11) What is your biggest complaint about smart speakers? (Biggest complaint)

2. Binary questions (Positive Thinking Scale)

(1) When somebody does something for me, I usually wonder if they have an ulterior motive. (N)
(2) I know the world has problems, but it seems like a wonderful place anyway. (P)
(3) I see much beauty around me. (P)
(4) I believe in the good qualities of other people. (P)
(5) When something bad happens, I ruminate on it for a long time. (N)
(6) When I see others prosper, it makes me feel bad about myself. (N)
(7) I think of myself as a person with many strengths. (P)
(8) When good things happen, I wonder if they will soon turn sour. (N)
(9) I savor memories of pleasant past times. (P)
(10) I regret many things from my past. (N)
(11) When I think of the past, for some reason the bad things stand out. (N)

`3. 5-point # questions (BREQ-3)`

(1) I take part in exercise because my friends/family/partner say I should.
(2) I consider exercise consistent with my values.
(3) I would feel bad about myself if I was not making time to exercise.
(4) I think it is important to make the effort to exercise regularly.
(5) I feel like a failure when I haven't exercised in a while.
(6) I think exercising is a waste of time.
(7) I consider exercise part of my identity.
(8) I find exercise a pleasurable activity.
(9) I can't see why I should bother exercising.
(10) I get restless if I don't exercise regularly.
(11) I get pleasure and satisfaction from participating in exercise.
(12) I feel under pressure from my friends/family to exercise.

`4. 5-point W questions (PUMP)`

(1) I need more time using my smartphone to feel satisfied than I used to need. (Tolerance)
(2) It would be very difficult, emotionally, to give up my smartphone. (Withdrawal)
(3) I have thought in the past that it is not normal to spend as much time using a smartphone as I do. (Longer time than intended)
(4) People tell me I spend too much time using my smartphone. (Great deal of time spent)
(5) I feel anxious if I have not received a call or message in some time. (Craving)
(6) I have used my smartphone when I knew I should be doing work/schoolwork. (Activities given up or reduced)
(7) When I stop using my smartphone because it is interfering with my life, I usually return to it. (Use despite physical or psychological problems)
(8) At times, I find myself using my smartphone instead of spending time with people who are important to me and want to spend time with me. (Failure to fulfill role obligations)
(9) I have almost caused an accident because of my smartphone use. (Use in physically hazardous situations)
(10) I have continued to use my smartphone even when someone asked me to stop. (Use despite social or interpersonal problems)

`5. 7-point # questions (FSQ)`

(1) I now would do anything to try to avoid a spider.
(2) Spiders are one of my worst fears.
(3) If I came across a spider now, I would leave the room.
(4) Currently, I sometimes think about getting bit by a spider.

(5) If I saw a spider now, I would ask someone else to kill it.
(6) If I encountered a spider now, it would take a long time to get it out of my mind.
(7) If I saw a spider now, I would feel very panicky.
(8) If I saw a spider now, I would probably break out in a sweat and my heart would beat faster.
(9) If I saw a spider now, I would think it will try to jump on me.

A.2 Questionnaires administered on the smart speaker

`1. Demographic questions`

(1) What is your age? (Age)
(2) What is your gender? (Gender)
(3) What is the highest degree or level of school you have completed? (Education)
(4) What is your current employment status? (Employment)
(5) Which languages are you capable of speaking fluently? (Language)
(6) How many smart speakers do you own? (Number of smart speakers)
(7) How did you obtain your smart speakers? (Ways of getting speakers)
(8) How long have you been using smart speakers? (Length of usages)
(9) How often do you use your smart speakers? (Frequency of usages)
(10) What do you enjoy most about smart speakers? (Enjoy most)
(11) What is your biggest complaint about smart speakers? (Biggest complaint)

`2. Binary questions (Positive Thinking Scale)`

(1) I am optimistic about my future. (P)
(2) I frequently compare myself to others. (N)
(3) I see my community as a place full of problems. (N)
(4) I see the good in most people. (P)
(5) When something bad happens, I often see a "silver lining," something good in the bad event. (P)
(6) When I see others prosper, even strangers, I am happy for them. (P)
(7) When I think of myself, I think of many shortcomings. (N)
(8) When good things happen, I wonder if they might have been even better. (N)
(9) When I think of the past, the happy times are most salient to me. (P)
(10) I think frequently about opportunities that I missed. (N)
(11) I sometimes think about how fortunate I have been in life. (P)

`3. 5-point # questions (BREQ-3)`

(1) I exercise because it is consistent with my life goals.
(2) It's important to me to exercise regularly.
(3) I can't see why I should bother exercising.
(4) I value the benefits of exercise.
(5) I enjoy my exercise sessions.
(6) I exercise because other people say I should.
(7) I exercise because it's fun.
(8) I feel guilty when I don't exercise.
(9) I exercise because others will not be pleased with me if I don't.
(10) I feel ashamed when I miss an exercise session.
(11) I consider exercise a fundamental part of who I am.
(12) I don't see why I should have to exercise.

`4. 5-point W questions (PUMP)`

(1) When I decrease the amount of time spent using my smartphone I feel less satisfied. (Tolerance)

(2) When I stop using my smartphone, I get moody and irritable. (Withdrawal)
(3) The amount of time I spend using my smartphone keeps me from doing other important work. (Longer time than intended)
(4) I think I might be spending too much time using my smartphone. (Great deal of times spent)
(5) When I am not using my smartphone, I am thinking about using it or planning the next time I can use it. (Craving)
(6) I have ignored the people I'm with in order to use my smartphone. (Activities given up or reduced)
(7) I have used my smartphone when I knew I should be sleeping. (Use despite physical or psychological problems)
(8) I have gotten into trouble at work or school because of my smartphone use. (Failure to fulfill role obligations)
(9) I have used my smartphone when I knew it was dangerous to do so. (Use in physically hazardous situations)
(10) My smartphone use has caused me problems in a relationship. (Use despite social or interpersonal problems)

`5. 7-point # questions (FSQ)`

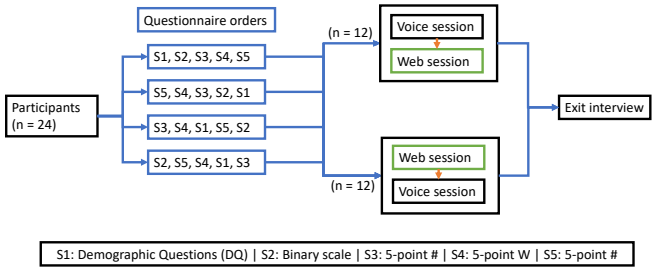
(1) Currently, I am sometimes on the look out for spiders.
(2) I now think a lot about spiders.
(3) I would be somewhat afraid to enter a room now, where I have seen a spider before.
(4) If I encountered a spider now, I would have images of it trying to get me.
(5) If I came across a spider now, I would get help from someone else to remove it.
(6) If I encountered a spider now, I wouldn't be able to deal effectively with it.
(7) If I saw a spider now, I would be afraid of it.
(8) I would feel very nervous if I saw a spider now.
(9) If I saw a spider now, I would think it will harm me.

A.3 Voice commands for voice surveys

(1) Hey Google, talk to monkey forecast. → Demographic questions
(2) Hey Google, talk to monkey forecast about positive thinking questions. → Binary questions
(3) Hey Google, talk to monkey forecast about number activity survey. → 5-point # questions
(4) Hey Google, talk to monkey forecast about smartphone survey. → 5-point W questions
(5) Hey Google, talk to monkey forecast about spiders. → 7-point # questions

A.4 Counterbalancing survey orders

Here, we show the 8 counterbalancing experiment procedures:



Received January 2022; revised April 2022; accepted August 2022