

Screen or No Screen? Lessons Learnt from a Real-World Deployment Study of Using Voice Assistants With and Without Touchscreen for Older Adults

Chen Chen
Computer Science and Engineering
University of California San Diego
La Jolla, CA, United States
chenchen@ucsd.edu

Ella T. Lifset
Biological Sciences
University of California San Diego
La Jolla, CA, United States
etlifset@ucsd.edu

Yichen Han*
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, United States
yichenha@andrew.cmu.edu

Arkajyoti Roy
Department of Mathematics
University of California San Diego
La Jolla, CA, United States
aroy@ucsd.edu

Michael Hogarth
School of Medicine
University of California San Diego
La Jolla, CA, United States
mihogarth@ucsd.edu

Alison A. Moore
School of Medicine
University of California San Diego
La Jolla, CA, United States
alm123@ucsd.edu

Emilia Farcas
Qualcomm Institute
University of California San Diego
La Jolla, CA, United States
efarcas@ucsd.edu

Nadir Weibel
Computer Science and Engineering
University of California San Diego
La Jolla, CA, United States
weibel@ucsd.edu

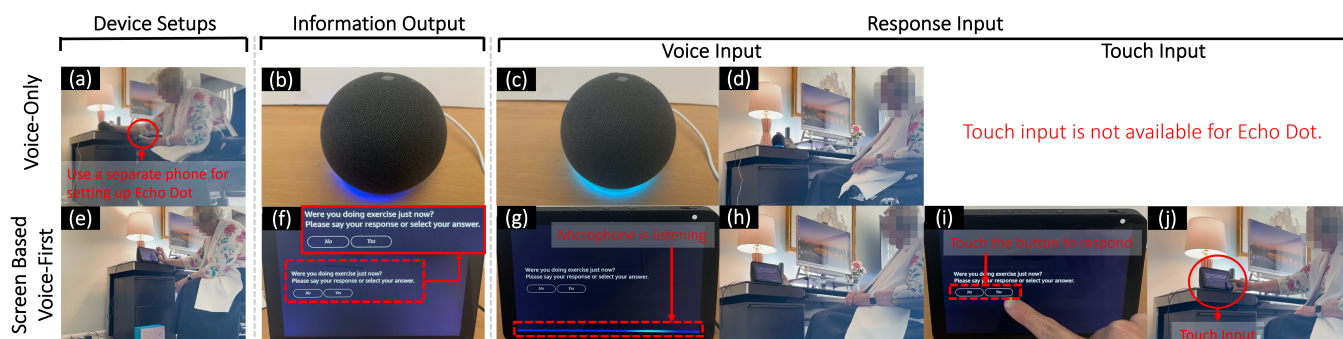


Figure 1: Our real-world deployment study aims to understand the affordances brought by the built-in touchscreen of Voice Assistants (VAs). We focus on device setup phases (a, e), as well as long-term uses for conducting diary survey and other miscellaneous general purposes. Notably, the touchscreen-based voice-first VAs allow older adults to *see* prompts (f) and input responses by either *voice* (g - h) or *touch* (i - j), compared to the voice-only VAs (b - d).

ABSTRACT

While voice user interfaces offer increased accessibility due to hands-free and eyes-free interactions, older adults often have challenges such as constructing structured requests and perceiving how such devices operate. Voice-first user interfaces have the potential

to address these challenges by enabling multimodal interactions. Standalone voice + touchscreen Voice Assistants (VAs), such as Echo Show, are specific types of devices that adopt such interfaces and are gaining popularity. However, the affordances of the additional touchscreen for older adults are unknown. Through a 40-day real-world deployment with older adults living independently, we present a within-subjects study ($N = 16$; age $M = 82.5$, $SD = 7.77$, $min. = 70$, $max. = 97$) to understand how a built-in touchscreen might benefit older adults during device setup, conducting self-report diary survey, and general uses. We found that while participants appreciated the visual outputs, they still preferred to respond via speech instead of touch. We identified six design implications that can inform future innovations of senior-friendly VAs for managing healthcare and improving quality of life.

*The author contributed to the project while at the University of California San Diego.



This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

ASSETS '23, October 22–25, 2023, New York, NY, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0220-4/23/10.
<https://doi.org/10.1145/3597638.3608378>

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Field studies*.

KEYWORDS

Older Adults, Voice Assistants (VAs), Real-World Deployment Study

ACM Reference Format:

Chen Chen, Ella T. Lifset, Yichen Han, Arkajyoti Roy, Michael Hogarth, Alison A. Moore, Emilia Farcas, and Nadir Weibel. 2023. Screen or No Screen? Lessons Learnt from a Real-World Deployment Study of Using Voice Assistants With and Without Touchscreen for Older Adults. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*, October 22–25, 2023, New York, NY, USA. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3597638.3608378>

1 INTRODUCTION

Voice user interfaces offer increased accessibility due to the nature of hands-free and eyes-free interactions [32, 41, 59]. However, older adults often have challenges such as constructing structured requests and perceiving how such devices operate [42]. Touchscreen-based voice-first user interfaces, referring to those that “*primarily accept user input via voice commands, and may augment audio output with a tightly integrated screen display*” [71, 77], have the potential to address the aforementioned challenges by enabling multimodal interactions [28]. Standalone voice + touchscreen Voice Assistants (VAs) (e.g., Echo Show [36]) are specific types of devices that adopt such interfaces and are gaining popularity among young people [28]. However, the affordances [55] of the additional touchscreen — both in terms of helping users and preventing them from achieving their goals — are still unknown for older adults. For example, it is unclear if and how older adults appreciate the affordances of the touchscreen of a device like the Echo Show [36] (Fig. 1e - j) in comparison to its voice-only counterpart (i.e., Echo Dot [35], Fig. 1a - d); the Echo Show allows users to input commands through *touch* or *speak*, and *see* additional visual elements along with the audio response, yet it also brings setbacks such as larger form-factor and more complicated visual interfaces.

Existing research has explored how older adults have used and perceived existing features of voice-only VAs [16, 43, 57, 58, 72], but the potential merits and setbacks of the secondary touchscreen modality are underexplored. Additionally, most prior works (e.g., [57, 72]) only focused on general uses (e.g., music) of VAs for older adults, rather than healthcare applications, which are indispensable use cases [15, 16].

We present a real-world within-subjects study ($N = 16$; age $M = 82.5$, $SD = 7.77$, $min. = 70$, $max. = 97$) to understand the affordances that the additional touchscreen for standalone VAs could bring to aging populations. Besides general uses (e.g., [43, 72]), we investigate the feasibility of using voice to collect self-report End-of-Day (EOD) diary survey (hereafter referred to as *diary*) [69], which help clinicians better understand older adults’ life routines and healthcare needs. Our study, based on Echo Dot and Show, aims to address three key Research Questions (RQs):

- **(RQ1)** How does the built-in touchscreen affect the older adults’ experience of *setting up the device*?

- **(RQ2)** How does the built-in touchscreen affect the older adults’ behaviors and experience of *conducting self-report daily diary survey*?
- **(RQ3)** How does the built-in touchscreen affect the older adults’ behaviors and experience of using VAs for *general purposes*?

In collaboration with UC San Diego Health and the Vi at La Jolla, we deployed both devices in real-world older adults’ residences for a total of 40 days. Our findings include three aspects: **(1)** During device setup, older adults appreciated the merits of the additional touchscreen. Quantitatively, we measured an overall reduction in the task completion time of around 50% while setting up the Echo Show, compared to the counterpart; **(2)** While conducting self-report diary survey, participants suggested that both systems needed to be more interactive and conversational. But overall, they enjoyed the visual output enabled by the additional display. Despite this, participants still responded more often to survey questions via speech over touch, although the time needed for touch input was characterized by an approximately 20% reduction; **(3)** For general uses, older adults appreciated the visual outputs and acknowledged the sense of companionship with the voice-first modality. The additional visual information (e.g., visual texts and icons) also encouraged older adults to engage more with the VAs. However, also in this case, interactions were mostly based on speech input instead of touch. Based on these findings, we identified six design implications under two themes that can inform future innovations of senior-friendly VAs for managing healthcare and improving Quality of Life (QOL). We believe our work will impact practitioners and researchers attempting to design senior-friendly voice-first VAs for aging populations, both for enhancing their healthcare and to better support general uses.

2 RELATED WORK

2.1 Design of Voice-Based Virtual Assistants (VAs)

VAs refer to software agents that listen and respond to verbal commands [20] and have been integrated into a heterogeneous hardware embodiment. Fig. 2 shows a taxonomy based on two dimensions: *modality* (i.e., voice-only or voice + touchscreen) and *the way the devices are incorporated into users’ lives* (i.e., user-attached or -detached).

Voice vs. Voice + Touchscreen. The simplest VA embodiment is the smart speaker, where voice is the *only* supported modality for both queries and responses. While offering hands-free and eyes-free control, voice-only interfaces present two major issues. First, receiving information only by voice is often ambiguous and inefficient for information consumption due to sequential information access, in contrast to the visual scanning [77]. Second, voice-based interactions yield more turn-taking compared to screen-based interactions [60]. To address these, existing research proposed the concept of touchscreen-based voice-first user interfaces, referring to those VAs whose primary functionality can be accessed through speech, but have a touchscreen as an auxiliary medium for information input and output [3, 71]. Such interfaces bring together the merits of voice (an efficient input modality) and screen (an efficient output modality) [77]. Furthermore, recent works (e.g., [11, 51])



Figure 2: Design taxonomy of VAs. On one extreme, the *user-attached* and *voice + touchscreen* design provides the highest interactivity, yet might introduce complexity during operation, setup, and troubleshooting. On the other extreme, the *user-detached* (or *standalone*) and *voice-only* design offers the simplest way for device uses, yet provides the lowest interactivity.

demonstrated that the embodied conversational agents – VAs that leverage the display to show a visual representation of a human [5] – are preferred by older adults in terms of social isolation and loneliness.

User-Attached vs. User-Detached devices. User-attached VAs are voice-enabled devices that need to be worn or handheld by users (e.g., Siri on iPhone), and are involved with repetitive maintenance related tasks (e.g., charging devices). In contrast, standalone VAs (e.g., smart speakers and displays) usually need an external power supply and are expected to be placed in a particular environment permanently. These devices are promising for older adults, as they only need to be set up once and can run continuously, so users can focus on task-related operations (e.g., making queries) [16]. Our study only involved user-detached (*i.e.*, standalone) devices.

2.2 Usability Evaluation and Real-World Deployment of Voice User Interfaces for Older Adults

While existing research recognized the promise of using voice for information interactions for older adults [16, 68], most work focused on evaluating *voice-only* VAs from the older adults’ perspective. For example, Jesús-Azabal *et al.* [39] introduced *Remembranza*, a medication reminder skill based on Echo speakers. Trajkova *et al.* [72] investigated older adults’ uses of smart speakers, and found that most participants become non-users due to the lack of perceived usefulness. Upadhyay *et al.* [74] studied the explorations and long-term uses of VAs by older adults living in a long-term care community, however the affordances of touchscreen were not explored. Kakera *et al.* [40] identified the usefulness of multiple features that can be realized by VAs for supporting older adults living independently, yet the voice-first VAs were excluded. Choi *et al.* [19] suggested the most frequently used features of Echo speaker by older adults are asking practical questions and managing tasks. Few recent works

[18, 38, 72] emphasized that although voice-only VAs do not require older adults to be technology savvy, the low interactivity might fail to support older adults’ needs, especially when it comes to the management of health data. Similarly, Nallam *et al.* [54] suggested that older adults see the potentials for using VAs to search for health information and support health tasks, yet adoptions of VAs could be affected by access barriers, confidentiality risks, and receiving trusted information. Pradhan *et al.* [57] investigated how older adults treat the smart speakers as a human. They also explored how Echo Dot could be used by older adults with low technology experience [58]. Bonilla *et al.* [8] explored older adults’ understanding of VAs’ privacy and security implications. Kim *et al.* [43] conducted a longitudinal study to understand older adults’ perception and use of Google Mini [70]. They also explored the initial interactions of older adults while using a smart speaker. Shade *et al.* [64] focused on medication reminders using Google Home Mini.

Researchers also investigated heterogeneous multi-modal voice-first interfaces to help improve older adults’ QOL. Shalini *et al.* [65] designed a system for older adults to track health information (e.g., sleep quality) inferred by instrumented in-home sensors, using both audio and visual display. Barros *et al.* [4] conducted a usability assessment of smartphone Google Assistant and Siri, and found that users prefer the Siri interface because it is minimalist. Hu *et al.* [34] showed the designs of seven types of speech acts for older adults by leveraging the built-in touchscreen using politeness theory. Gustafson *et al.* [30] showed that using touchscreen-based VAs for delivering eHealth interventions is more effective compared to using laptops. Further, researchers also investigated the integration of VAs with heterogeneous smart home devices. For example, Kowalski *et al.* [46] studied how VAs benefit older adults when they are integrated with smart home technologies. Ennis *et al.* [26] incorporated Echo into smart cabinets to support older adults’ independence. Valera Román *et al.* [75] studied how the combination of smart bracelets, smart home devices, and VAs could allow older adults to monitor their physical activity and sedentary patterns.

In contrast to existing research that only focuses on voice-only VAs and/or the techniques to design touchscreen-based VAs for specific types of interactions, we explored *how the voice + touchscreen VAs could influence older adults’ experience of device setups, conducting self-report diaries, and general uses* through a real-world deployment.

2.3 Ecological Momentary Assessments (EMA) and End-Of-Day (EOD) Diaries Data Collections

Ecological Momentary Assessment (EMA) involves repeated sampling of subjects’ current behaviors and experiences in real time in their environment [67]. EMA can assist medical providers to better understand patients’ daily routines and healthcare needs, particularly important for older adults with chronic diseases [16]. Moskowitz *et al.* [53] classified EMA into three types: *diaries* (fixed interval assessment with a frequency of once per day, employing retrospective coverage strategy [67]), *experience sampling* (using specific signaling devices that randomly notify participant to make

ID	Age	Sex	Education	Occupation Before Retirement	Past Experience "I am familiar with VAs"	TechPH		MDPQ	CPQ	Significant Health Condition that might Hinder Using VAs
						Enthusiasm	Anxiety			
P1	90 - 95	F	Doctorate Degree	Social Sciences Researcher	Somewhat Agree (4)	3.3	3.3	3.3	3.9	Significant hand tremors
P2	80 - 85	F	Professional Degree	Healthcare Worker	Strongly Agree (5)	4.7	2.3	4.8	4.8	N/A
P3	70 - 75	F	Professional Degree	Administrator	Somewhat Disagree (2)	3.7	4.7	2.4	4.3	N/A
P4	70 - 75	F	Professional Degree	Administrator	Strongly Agree (5)	3.3	4.7	4.3	4.4	N/A
P5	90 - 95	M	Professional Degree	Healthcare Worker	Somewhat Agree (4)	4.3	2.0	4.8	4.9	Hearing Impairment
P6	75 - 80	M	Some College	Businessman	Somewhat Agree (4)	4.7	3.3	4.4	4.5	N/A
P7	75 - 80	F	Professional Degree	Social Worker	Somewhat Agree (4)	2.0	4.3	5.0	4.0	N/A
P8	75 - 80	M	Professional Degree	Consultant	Somewhat Disagree (2)	4.3	3.0	1.0	4.3	N/A
P9	75 - 80	F	Professional Degree	School Teacher	Strongly Agree (5)	5.0	3.3	4.5	4.6	Chronic back pain
P10	75 - 80	M	Bachelor's Degree	Accountant	Somewhat Disagree (2)	3.0	3.7	4.0	4.3	N/A
P11	95-100	M	Doctorate Degree	Administrator	Somewhat Disagree (2)	3.0	2.7	4.4	4.6	Hearing, speech, and mobility impairment
P12	75 - 80	M	Bachelor's Degree	Writer	Somewhat Disagree (2)	3.7	4.3	3.1	4.3	Mobility impairment
P13	75 - 80	M	Doctorate Degree	College Professor	Somewhat Disagree (2)	3.3	3.3	4.8	4.8	N/A
P14	85 - 90	F	Bachelor's Degree	Administrator	Strongly Agree (5)	3.7	3.7	4.4	4.8	N/A
P15	90 - 95	M	Doctorate Degree	College Professor	Strongly Agree (5)	5.0	3.7	4.8	4.6	N/A
P16	85 - 90	F	Professional Degree	Social Scientist	Neither Agree Or Disagree (3)	3.3	5.0	2.9	3.6	N/A
<i>M (SD)</i> 82.5 (7.77)						3.8 (0.81)	3.6 (0.84)	3.9 (1.07)	4.4 (0.35)	

Figure 3: Participants' demographics. All self-reported scores are on the scale of 1 to 5. Participants proficient with technology would have high score of VA Past Experience, TechPH (Enthusiasm), MDPQ, and CPQ, and a low score of TechPH (Anxiety).

reports a fixed number of times per day [53]), and *event-based sampling* (self-reports are solicited at the time the variable of interest such as physical activity takes place [7]).

While online survey is a common and simple strategy for collecting EMA data (e.g., [22]), this method is usually limited in terms of compliance and accessibility. Some researchers also explored using smartphones to collect diaries (e.g., [1]) possibly because of the widely accessible of smartphones compared to computers. However, the low efficiency of typing due to finger dexterity problems [76] and the complexities of troubleshooting these devices might be problematic for older adults. The recent μ EMA [37] demonstrated the effectiveness of using smartwatches to conduct microinteraction-based event sampling, through which participants could answer questions by a quick tap on the smartwatch. However, typing on a smartwatch for open-ended questions is impractical.

Voice has been identified as a promising approach for collecting survey data from older adults [15, 16, 48]. Prior researchers have used interactive voice response systems to address the limitations of mobile and wearable devices [23]. Instead, we focus on using *standalone VAs* for *self-report diary* data collection, where participants need to retrospectively respond to a set of questions for the past 24 hours. We focus on standalone devices, for two reasons. First, unlike the young adults, older adults are more likely to be at home more often [14, 50]. This phenomenon was more prominent during the COVID-19 pandemic and the social distancing restrictions, particularly for older adults who are known to be a group at higher risk [27]. Second, we only focus on *diary* studies that are not time sensitive (Sec. 3.2 and Appendix A). Participants were expected to complete the survey on a daily basis, but the time did not have to be strictly specified.

3 METHODS

3.1 Participants and Study Procedures

We recruited 16 older adults, including eight males and eight females, through UC San Diego Health¹ and the Vi at La Jolla Village²

¹UC San Diego Health: <https://health.ucsd.edu> [Accessed on 7/1/2023]

²The Vi at La Jolla Village: <https://www.viliving.com/locations/ca/san-diego-la-jolla> [Accessed on 7/1/2023]

(age, $M = 82.5$, $SD = 7.77$, $min. = 70$, $max. = 97$, see Fig. 3). All participants are self-identified as capable of living independently. The study was approved by the Institutional Review Board (IRB), and the Echo Dot and Echo Show (~\$130) were provided as incentives. Our study was conducted during December to February, during which all participants resided at home during majority of the time without travel plans. Overall, our study was structured in four phases, lasting for 40 days at the residences of older adults:

Phase 1: Pre-Study Questionnaires. Participants completed informed consent and questionnaires, including their past experience using VAs, as well as three validated questionnaires investigating their attitudes towards technology and their experiences with it (see Fig. 3).

- **TechPH (Technophilia)** [2] measures the older adults' attitudes toward technology. We reported the average score for their *enthusiasm* and *anxiety* toward general technologies. A high enthusiasm and low anxiety imply a positive attitude as measured through this instrument.

- **MDPQ (Mobile Device Proficiency Questionnaire)** [56, 62] evaluates the proficiency with smartphones for older adults. While smartphones are not our focus, we extrapolated that the older adults would transfer existing skills from interactions with smartphones to voice-based interactions. A high MDPQ score indicates a highly proficient smartphone user.

- **CPQ (Computer Proficiency Questionnaire)** [9] evaluates the older adults' proficiency with desktop computers. Similar to the MDPQ, we hypothesized that older adults might transfer some existing computer skills to voice-based interactions. A high CPQ score indicates a highly proficient desktop computer user.

Notably, while focusing on VAs, the results of MDPQ and CPQ could reflect and imply older adults' skills, experience, and attitude toward using VAs from a broader context by looking at richer perspectives of technology exposures and uses in their life.

Phase 2: Device Setup. Our second phase of the study aims to address RQ1. Although VAs may often be set up by others, it is

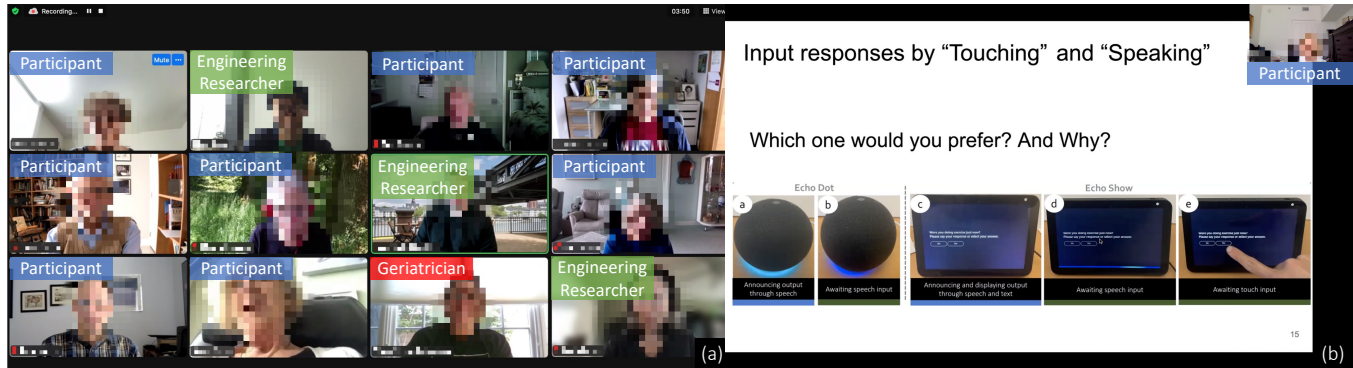


Figure 4: (a) One of the focus groups conducted through Zoom with eight participants, three engineering researchers, and one geriatrician; (b) One participant (top right corner) was explaining her ideas of Q6 of Fig. 15 in Appendix B.

still important for older adults to know how to initialize the device and finish the last-mile tasks (e.g., connect the device to the WiFi) [16, 17]. A *within-subject design* was used to evaluate participants' performance on setting up VAs with and without touchscreen. We pre-setup and initialized the devices with our Alexa skill, yet we required participants to engage in the last-mile tasks. An experimental Amazon account was created for each participant, enabling us to track participants' responses and usages. During in-person meetings, we first introduced the scope of our project, described and had participants signed all required forms, and explained the technology related to VAs. We then invited participants to set up two devices *independently*. Participants were asked to use the official instruction manuals as needed. Eight participants were prompted to setup Echo Dot first, followed by Echo Show. While another eight participants were prompted to setup Echo Show first, followed by Echo Dot. While setting up the devices, participants were required to type the username and password of our pre-created experimental account, with the average number of characters being 33 ($SD = 2.28$) mixed with letters, numbers and special characters. Assistance was provided, *if and only if* the participants gave up on the effort. Semi-structured interviews were conducted after participants set up each VA. The questions used to guide the interviews could be referred to Appendix B. While we excluded the quantitative measures from participants who gave up on setting up device(s), participants were still encouraged to discuss about their feeling after observing the research assistant(s) setting up the device(s) on their behalf. This phase on average spent 37.03 min ($SD = 9.50$ min) with each participant.

Phase 3: Investigations of VA Uses for Conducting Self-Report Diaries and General Uses. We used a *within-subject design* to evaluate the values of the additional touchscreen while using VA to conduct self-report diary survey (RQ2) and for general uses (RQ3). Eight participants were asked to use Echo Dot first, followed by Echo Show, each for 15 days, while another eight participants were instructed to use Echo Show first, followed by Echo Dot. Before each 15-day session, each participant was instructed on how to use VAs with official user manuals and provided with five days to explore and familiarize themselves with the given devices. During each 15-day session, participants were instructed to use the features of their VAs for general uses for added convenience and benefit to their

daily routine. Additionally, since we aimed to explore the feasibility and usability of using VAs as a tool for conducting self-report daily diaries, we designed a set of diary questions for wellness screening (Sec. 3.2). Participants were expected to complete the diary on a daily basis, but were not required to complete it at specific time of the day, and could choose when to engage with the device. At the end of the diary survey, a usability question was delivered to each participant, using the prompt: “on a scale of 1 – 5, how do you like to use voice assistant to report your diary survey? 1 being dislike extremely and 5 being like extremely”. This offered us insights on the participants' overall experience after each time they used VAs for the diary survey. Participants were encouraged to choose their preferred methods to remind themselves. At the end of each 15-day session, participants were invited to rate how strongly they agreed with the following three statements in a 5-point Likert scale:

- (Q1) “Conducting the daily diary using the given VA could cause interruption burden to my daily life routine”;
- (Q2) “I am comfortable to use the VA for general uses”;
- (Q3) “It is easy to use the VA for general uses”;

Participants were then invited to complete NASA TLX [33] and System Usability Scale (SUS) [12, 13] questionnaires regarding their overall user experience. To minimize the time and effort needed to complete the questionnaires, we excluded the pair-wise workloads comparisons in TLX, and assumed the weights for each perceived workloads were identical while computing the overall TLX score. We then conducted a remote semi-structured interview with each participant. All procedures were repeated for the second 15-day session with the other device.

Phase 4: Focus Groups. We adopted Robson *et al.*'s suggestion [61] regarding the size of the focus group to be eight to twelve for an in-depth discussion, and therefore organized two online focus groups (Fig. 4). The attendees of each focus group included eight participants, three engineering researchers, and one geriatrician. Same prompts (Fig. 15 in Appendix B) and slides were used to guide the discussion in both focus groups.

3.2 Design of the Diary Survey

To understand the affordances of the touchscreen while using standalone VAs to collect diaries, we designed a self-report diary survey with geriatricians from the *Anonymous Academic Medical Center*

and empirical guidance from the World Health Organization [78]. Our diary survey was centered around the eight themes: quality of sleep, social interactions, exercise, pain management, alcohol uses, food consumption, symptoms, and medication management. We also added a set of usability questions for the purpose of our study (Appendix A). While establishing the validity of diary data acquired is beyond our scope, we have iteratively designed questions with one geriatrician that could be answered easily and quickly (the whole survey would typically take ≤ 5 min and could be interrupted at any time, verified during pilot testing).

We expected four types of answers: *binary*, *Likert*, *number*, and *open*. The design of the first three types of responses has been widely used in many existing EMA studies (e.g., [24, 37]). Participants were asked to speak their choice, and were provided with the additional alternative option to input response by *touching* the buttons on the touchscreen while using the Echo Show (Fig. 1i). We adopted the same design as μ EMA [37], and provided five options (i.e., five buttons) for *Likert*- and *number*-type responses with touch input. We designed the height of each button to be approximately the same as the finger width, and the width of the button being around twice the finger width (Fig. 1i). This decision was made to ensure all buttons are easy to be clicked while fitting on the same screen. Following the suggestion from the geriatrician, we also included an *open* unstructured type of response that supported only the input modality via speech. We did not include healthcare feedback to the participants' reported data due to liability concerns from our institution's IRB. Our study only focused on data collection and observation of behavior, not intervention to modify behavior.

3.3 Implementation

We selected Echo Dot [35] (Fig. 1a - d) and Echo Show with a built-in eight inches touchscreen [36] (Fig. 1e - j) as the testbed due to the dominant market share [44]. However, the majority of our findings could be transferred to other similar standalone VAs. We implemented an Alexa skill and a Flask backend to track the conversation states. "My Health" was used as the invocation name, which could be easily remembered and clearly spoken as verified during our pilot testing.

3.4 Measures and Data Analysis

We structured our analysis based on the experience of device setup (RQ1), conducting self-report daily diaries (RQ2), and general uses (RQ3). We describe the measures and approaches of analyzing data for each aspects.

Device Setup. By observing participants and analyzing official instruction manuals, we first summarized six key steps (Fig. 5a) while setting up the Echo Dot and Echo Show. While observing participants' behavior of setting up devices, we noted the Task Completion Time (TCT) that participants spent during each step, which were then be used to compute the overall TCT. Notably, we used the *accumulated* time spent on each step as the final TCT for analysis purposes, because, for example, the participant might read instructions after every setup actions interleavedly. We did not include other actions except the six pre-defined steps in the overall TCT (such as finding the WiFi password or handling unexpected phone calls), since they are not related to the assigned tasks and vary greatly

among participants. To analyze the data quantitatively, we used Repeated Measure Analysis of Variance (RM-ANOVA) ($\alpha = .05$) to evaluate the statistical significance of the effects of the two devices. The Tukey's Honestly Significant Difference (HSD) test [73] was used for conducting *post-hoc* test. Before performing RM-ANOVA, we first conducted the normality check of measures in each catalogue using Shapiro-Wilk test [66]. For those failing to pass the normality check, we adopted Aligned Rank Transform (ART) [79] for statistical significance test, followed by ART-C with Bonferroni adjustment³ [6, 25]. For all statistical significance analysis, the partial eta square (η_p^2) was used to understand the effect size, with .01, .06, and .14 being used as the empirical thresholds for small, medium and large effect sizes, where a larger effect size indicates a higher practical significance [21]. Two researchers then performed thematic analysis [10] and adopted a mixture of emergent and priori coding approaches on the 9.87 hours of video-audio recordings. Specifically, we first transcribed the recorded audio clips and removed the connecting phrases (e.g., "[...] you know [...]") to enhance readability. We then closely read the transcripts and watched the recorded videos iteratively, and allowing codes to emerge freely from the data. During analysis, we held multiple discussions to discuss and iteratively refine the codes and reconcile the disagreement. Overall, four iterations were conducted to ensure the reliability of coding results. Fig. 16 in Appendix C shows the codebook.

Experience of Conducting Self-Report Daily Diaries. We analyzed interaction traces related to self-report diary reportings, where each *interaction trace* refers to a pair of request and response. We adopted four quantitative measures that were introduced in [37], including:

- **Diary Survey Compliance Rate**, defined by the percentage of surveys being *fully* completed, versus the number of surveys that were expected to be completed. A failure of survey compliance could be caused by either forgetting to start the survey, or failing to complete all designated questions by stopping mid survey. Overall, survey compliance measures the performance of the diary data collection tool on the survey level (i.e., question set);
- **Question Completion Rate**, defined by the percentage of the questions being answered over the total number of questions being delivered. Unlike survey compliance, question completion measures the feasibility of the dairy data collection tool on the individual question level;
- **Initial Prompt Response Rate**, defined by the percentage of questions completed when delivered the first time. If the participant does not answer as expected (such as wrong format or without speaking any content), the system repeats the question;
- **Response Latency**, measured by the elapsed time in milliseconds (ms) between the instant when a specific question is announced and the response for that particular question has been input, either by *voice* or *touch*;

A similar method as device setup was then used to analyze the quantitative measures of aforementioned measures. While performing statistical significance analysis of response latency related to

³For the statistical analysis by ART, we reported the degree of freedoms of F -statistic of the aligned and ranked responses instead of original observations [79]. We used ARTool (<https://depts.washington.edu/acelab/proj/art/>) [Accessed on 7/1/2023] for conducting this analysis.

touch input, we excluded participants who did not use the touch input over the course of using Echo Show. With the same approach for analyzing qualitative data collected during device setups, we evaluated the qualitative data collected after each of 15-day session (around a total of eight hours interviews over the phone), as well as the focus groups (around a total of two hours video-audio recordings). Fig. 17 in Appendix C shows the codebook.

Experience of Using General Features. With the participants' consent, we downloaded all the interaction logs captured by Amazon Privacy Portal⁴ and analyzed a total of 4350 requests related to general uses. We manually inspected the logs and identified *interaction sessions* to evaluate the participants' uses of built-in features, where one interaction session contains all requests and responses when a specific skill/feature is used. Unlike Kim *et al.* [43], who used the pairs of request-response communications, interaction sessions provide a better quantitative measurements of features in use. This is because some features (e.g., chat and knowledge query) will inherently introduce more follow-up questions compared to others (e.g., service), and therefore considering request-response pairs in isolation do not reflect usage frequencies of particular features. Notably, the logs generated during initial five-day training sessions were not included in our analysis. We first carefully read through all requests and responses, and used an emergent coding approach to tag each interaction session. Similar to the qualitative analysis of interview data, three researchers analyzed the logs and discussed to refine the codes iteratively and reconcile disagreements. To quantitatively understand participants' usages, we then used the same method as device setups to evaluate the statistical significance of the frequency of the captured interaction session for each theme over the interface type, the overall NASA TLX, and SUS responses. Semi-structured interview data collected after each 15-day session and the focus groups related to general uses were then analyzed using the same method as device setups. Fig. 17 in Appendix C shows the codebook.

4 RESULTS

Our results are organized based on the three RQs, which aim to investigate the impacts of touchscreen during (RQ1) device setup (Sec. 4.1), (RQ2) conducting self-report diary survey (Sec. 4.2), and (RQ3) general uses (Sec. 4.3).

4.1 RQ1: How Does the Built-In Touchscreen Affect the Older Adults' Experience of Setting Up Devices?

Overall, although most participants felt it was a daunting task to set up the designated devices (e.g., "it was a little bit scary because there are a bunch of buttons and things to be pressed" (P13)), 11 participants were able to fully set up both devices. Specifically, P4 gave up during the Echo Show setup phases due to unexpected personal duties; P6 gave up on typing the login credentials for Dot; P7 gave up setting up both devices due to the lack of interest and confidence; P11 and P12 gave up on setting up both devices due to the inconveniences caused by impaired mobility. Our results show that participants

took significantly less TCT to set up the Echo Show than Echo Dot ($F_{1,20} = 8.57$, $p = .003$, $\eta^2 = 0.37$, Fig. 5h). Through qualitative analysis, we now discuss potential reasons and participants' user experience.

Using the standalone touchscreen could enhance typing experience. Due to the lack of touchscreen of Echo Dot and the need to input credentials during the signing in phase, participants adopted different strategies to type on mobile devices, such as simply using fingers and the on-screen virtual keyboard (P3, Fig. 6a), the stylus (P6, Fig. 6b), or adopting a tablet with bigger display and external keyboard (P8, Fig. 6c). Most participants recognized the merits of the built-in touchscreen for the enhanced typing experience. This has been validated by our measurements where a significant reduction of TCT for the signing in stage was observed while using the Echo Show ($F_{1,20} = 18.57$, $p < .005$, $\eta^2 = 0.48$, Fig. 5e). In contrast, no statistical significance were observed for reading instructions ($p = .543$), hardware connections ($p = .465$), and WiFi connections ($p = .214$). Participants' comments also reflected this observation. For example, P2 and P3 outlined how ease of typing is an important benefit: "inputting the data is the most helpful! because the screen was bigger than my phone." (P3) and "the underscore sign is a little bit hard to find on this phone." (P2). P10 emphasized the merits of on-screen keyboard: "the keyboard was very different! I preferred the [the virtual keyboard on the Show], as the bigger buttons are easy to be pressed!" P3 emphasized the issues that the well-known "fat finger" problem creates [76]: "[with my phone] I made a lot of mistakes, because it was small. And I missed entering information with my fat fingers". On the contrary, few participants still preferred to type on their phone due to the familiarity of everyday's hand-held mobile devices: "typing on built-in display was hard, probably because I'm more familiar with my phone" (P13).

Immediate and in situ visual feedback and guidance on built-in touchscreen could help track setup steps. Most participants preferred the *visual feedback* enabled by the built-in touchscreen, to the prompts on a separate smartphone app. Participants highlighted the helpfulness of integrating all interaction components in one single entity, leading to a better device setting up workflow. For example: "the touchscreen made the experience more streamlined" (P9). Although P11, who needs a wheeled walker due to mobility impairment, believed that the voice based interaction should be sufficient for general uses after initial impression (e.g., "I don't need the visual, just the voice is fine"), he still preferred the built-in touchscreen after observing the research assistant setting up both devices: "you don't have to worry about connecting two devices. You're dealing with one device where you have both the visual and the sound together. Whereas with the Echo Dot, you need to 'plug in' a separate phone in order to get the visual!" Further, many participants highlighted the benefits of immediate feedback given by the Show, and the consequent reduced demands on users' working memory [47]: "the setup was easier on the Show. Because we could actually see what we were doing. Whereas [with Dot] you're only hearing it and seeing it on the phone" (P2); "[Echo Show] gives me the directions right on the screen, then it would be easier than me looking at my phone and transferring the information mentally" (P4); and "[Echo Show] is better, because I have known the visual gave me immediate confirm

⁴Amazon Privacy Portal: <https://www.amazon.com/alexa-privacy/apd/home> [Last accessed on 7/1/2023].

(a) Descriptions of the steps to setup VAs

Step	Description	Dot	Show
Read Instructions	Participant reads the instruction manual to understand the following tasks;	Y	Y
Hardware Connections	Participant plugs the power strip into mains and female jack socket on the device;	Y	Y
Find Alexa App	Participant needs to find the Alexa App on mobile devices (e.g., phone). The time for downloading apps are excluded if the Alexa App is not pre-installed;	Y	
Signing In	Participant needs to type the given username and password to sign into experimental account;	Y	Y
Hit Setup Button	Participant hits the setup button to force device into set-up mode;	Y	
Connect to WiFi	Participant needs to input WiFi SSID and credentials to connect the devices to the internet;	Y	Y

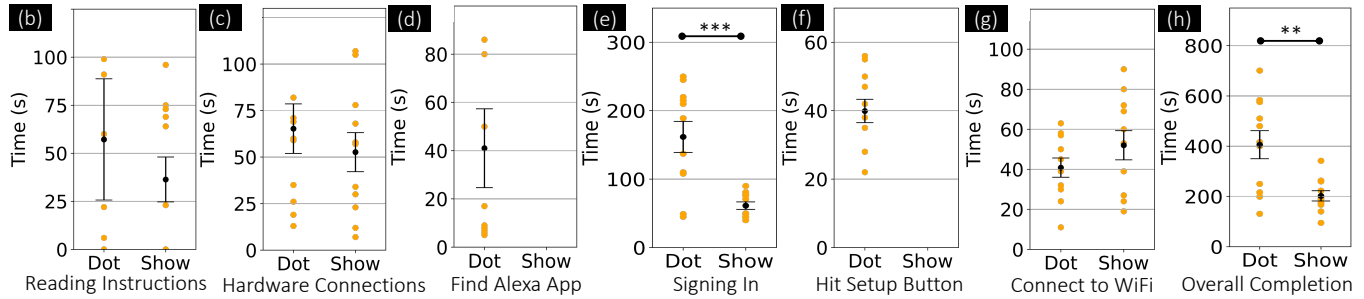


Figure 5: Evaluation results for participants setting up Echo Dot and Echo Show. (a) Description of steps for setting up VAs; (b) Task Completion Time (TCT) for reading instructions, (c) hardware connections, (d) finding the Alexa mobile app, (e) signing in with the experimental account, (f) finding and hitting the setup button, (g) connecting to WiFi, and (h) the overall TCT while setting up Echo Dot and Show. Notably, setting up Echo Show does not require participants to use the mobile Alexa app (d) and press the setup button (f). Participants (P4, P6, P7, P11, P12) who did not complete the tasks or gave up on the whole session were not included. We used standard error to represent the error bar. Notations for indicating the statistical significance of post-hoc test: * = .05 > $p \geq .01$, ** = .01 > $p \geq .001$, *** = $p < .001$.

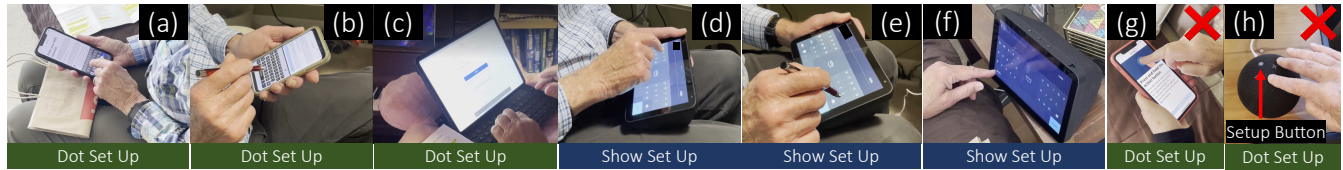


Figure 6: Typing on Dot (a - c) and Show (d - f). (a) and (d) show the methods adopted by the majority of users. (g) shows participants incorrectly consider the button image on the phone as the setup button of the Dot. (h) shows participants incorrectly think the mute button as the setup button.

of what I was doing” (P16). In contrast, the lack of direct and *in situ* visual guidance while setting up the Echo Dot could be one possible cause that led three participants to fail finding the setup button without hints. For the Dot, part of the instructions were on the Alexa phone app. When being instructed to “touch” the setup button, some participants (e.g., P1, Fig. 6g) incorrectly considered the button icon on the phone as the target, while others (P4, Fig. 6h) made incorrect attempts to interact with the mute button on the Dot. For example, P4 made this comment after being corrected by the research assistant: “I had assumed the button to push was the one on the top rather than the one on the side. [...] [the system should] tell me which button to push more precisely.”

The larger physical size of the touchscreen might be a hurdle. Without the need of a screen, the design of the Echo Dot is naturally smaller and lighter compared to Echo Show, and the merits of the small form factors of Echo Dot were outlined by half of participants. For example, P2 emphasized: “the Dot is smaller and more inconspicuous. So it’s easier to fit into a smaller space. I like the convenience of that [...] The Dot is less intrusive in your apartment

so that you can put it in different places more conveniently [...] I like the size of the Dot and the discreet shape of it”. After setting up the Echo Show, P12 commented: “[the Echo Show] would be too bulky! For me, you could see, I don’t have much space on the table. I have a small apartment, and I don’t have a lot of spaces or things. [...] For something that is new, you have to think about it, adjust it, and learn it”. Despite this, few participants suggested while the bulky size might affect the experience for setting up the Echo Show, it will not affect the long-term uses. For example, “the Echo Show is bulky, but I would just leave it there and don’t move it around” (P15).

4.2 RQ2: How Does the Built-In Touchscreen Affect the Older Adults’ Behaviors and Experience of Conducting Self-Report Daily Diary Survey?

Overall, participants demonstrated the usefulness of conducting such voice-first daily diaries (e.g., “it kind of reminded me to eat more fruit and many other aspects to keep myself healthy. So I thought it was helpful, just like a memory enforcement, as I don’t think it

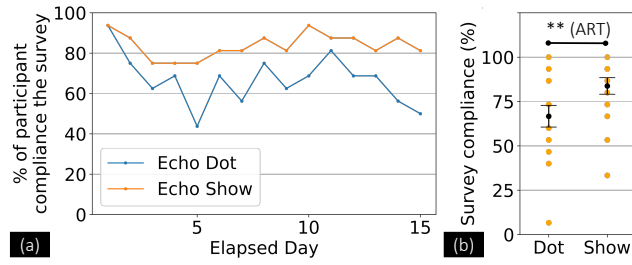


Figure 7: Evaluations of survey compliance. (a) % of participants that complied with the survey on each day; (b) Survey compliance rate on Echo Dot and Echo Show; Notations for indicating the statistical significance of *post-hoc* test: * = $.05 > p \geq .01$, ** = $.01 > p \geq .001$, *** = $p < .001$.

sometimes” (P14)). First, we demonstrate a statistical significance of the survey compliance rate (ART: $F_{1,15} = 10.33$, $p = .006$, $\eta_p^2 = .41$, Fig. 7b) while using the Echo Dot and Show. Fig. 7a further demonstrates the % of participants who complied with the diary over each of 15-day session, where participants using Echo Show exhibited a slight higher compliance rate from the second day, compared to the voice-only alternative. Second, while no statistical significance was detected in terms of question completion rate (ART: $p = .600$) and initial prompt response rate (ART: $p = .610$), a weak statistical increase of initial prompt response rate was captured (ART: $F_{1,15} = 5.12$, $p = .039$, $\eta_p^2 = .25$, Fig. 8a). Third, while using Echo Show for journaling diaries, participants adopted voice as the input modality, more frequent than that of touch input (ART: $F_{1,15} = 9.91$, $p = 0.007$, $\eta_p^2 = 0.40$, Fig. 8b), with an average of 73.06% versus 26.94%. Despite this, among 12 participants (except P3, P10, P11, and P15) who have used touch input, we found that using touch input enabled a shorter response latency significantly, compared to the voice counterpart (through Echo Dot or Show) (ART, $F_{2,34} = 14.40$, $p < .001$, $\eta_p^2 = 0.56$, Fig. 8c), with the average responses of using touch input being 6.056 seconds versus 7.639 seconds and 7.513 seconds using voice input by Echo Show and Echo Dot respectively. A similar observations have been captured in terms of measured response latency for *binary*- ($F_{2,33} = 7.08$, $p = .003$, $\eta_p^2 = .30$, Fig. 8d), *number*- (ART: $F_{2,16} = 5.16$, $p = .019$, $\eta_p^2 = .39$, Fig. 8e), and *Likert*- (ART: $F_{2,22} = 9.63$, $p < .001$, $\eta_p^2 = 0.47$, Fig. 8f) type questions. Fourth, we found a higher median of preference rating of using Echo Show to report diary survey compared to the voice-only counterpart by analyzing self-reported Likert scale of the usability prompt shown in Fig. 9a. Finally, Fig. 9b demonstrates that slightly more participants believed that using Echo Show will *not* cause interruption burden to their daily life, compared to using Echo Dot. Through qualitative analysis, we identified four findings.

Touch input is faster, but responding via voice is still preferred. Most participants appreciated the merits of hands-free interactions using voice to journal diaries (e.g., “it was interesting trying to [keep health diaries] when you weren’t sitting right at the device [the desktop PC] and it turned out that I had to do it using paper and pencil. Whichever devices or systems [VAs] that will help you do that would be very valuable” (P8)). However, nearly all participants subjectively believed that inputting response by touch could

be faster compared to using speech, which verified the validity of Fig. 8c. Some participants chose to use touch to interrupt the delivered prompts. Testimonies include: “with the Show, I don’t need to listen to the whole description. As soon as it is displayed, I know what to answer. I can move through the script faster” (P4), “having a touchscreen is faster. If I see something, I touch it [to submit my responses]; And it will go to the next question. If everything has to be oral, it has to ask me before I could answer. So with the touchscreen, it is a lot faster for things like going through checklists” (P13), and “response by touching speeds up quite a bit!” (P14).

However, participants overall responded more often to the prompts via speech than touch (73.06% vs. 26.94%), and this was confirmed by testimonies such as “I didn’t use the touchscreen for any general use, I just use the voice for interactions.” (P4). **Out of arms’ reach and inconveniences caused by impaired mobility** are the most common reasons discussed among participants. For example, “speaking is easier because you don’t have to lean over to press the buttons” (P1, with significant hand tremors) “if I am standing and I can reach [to the touchscreen]. I can respond quicker with a touch than I can with saying something and waiting for [the Echo Dot] to come back to me [through speech] with a question [...] that’s a lot faster for me to read than for me to listen to Alexa talking about it” (P10), and “usually I would be 10 feet away instead of have to be right next to it” (P12). P13 also appraised the voice over touch based on his past observations:

“If you can do it without walking over to it. That’s great! But if you have to walk over to it [to make a touch response], it can be a great hardship. [...] I was in the care center yesterday and there’s a guy who’s a 93 years old man who had surgery on his shoulder and his hip. He fell down and broke both his shoulder and his hip. He can’t get up and touch the screen. So for him, something that he could operate with just voice would be very important. So I think probably around 85 years old is when that starts to become an issue, the issue of get up and go, and touch the display, rather than interact verbally!” (P13)

This insights echos P9’s comment, who needs a wheeled walker for moving: “you have to be up close to the screen to really touch the button. Whereas, in terms of the Dot, you could be 15 or 20 feet away, and that’s not an issue”. Few participants also mentioned the **unpleasant visual experience due to discernible splotches caused by finger touch** For example, “when you have a touchscreen and touch it all the time it gets splotchy. So if there’s an alternative, like using your voice, sometimes I just simply prefer to do that.” (P2).

Responding by speech needs more support for controlling the conversation flow. While using speech as the input modality, all participants emphasized that the flow of conversation should be “more interactive and conversational” (P1). Participants identified the need for accepting longer responses (e.g., “I wanted to expand in an answer but there was no way of doing that” (P7)) and the short response time limit was compounded by the tendency of some participants to repeat the question at the beginning of their response. For example, without touchscreen, P8 sometimes exhibited the following interaction pattern, causing the system to fail in capturing valid diary responses:

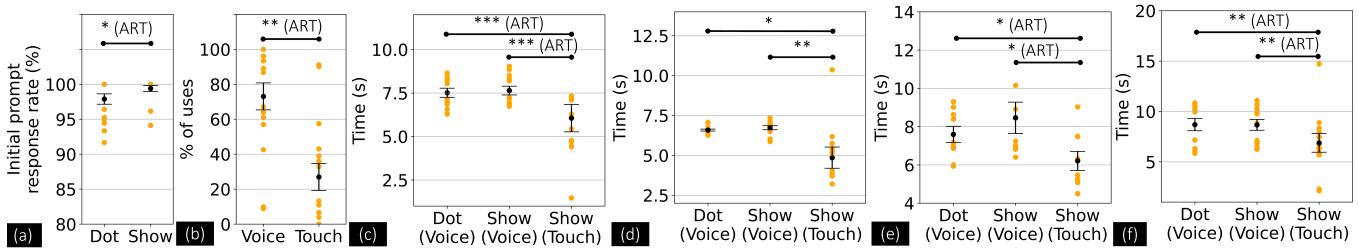


Figure 8: Evaluation results: (a) Initial prompt response rate for the *number*-type question, (b) % of uses of voice vs. touch while using Echo Show to complete the daily survey; (c) Overall response latency; (d - f) Response latency of *binary*- (d), *number*- (e) and *“Likert”*- (f) type questions. Notations for indicating the statistical significance of *post-hoc* test: * = .05 > $p \geq .01$, ** = .01 > $p \geq .001$, *** = $p < .001$.

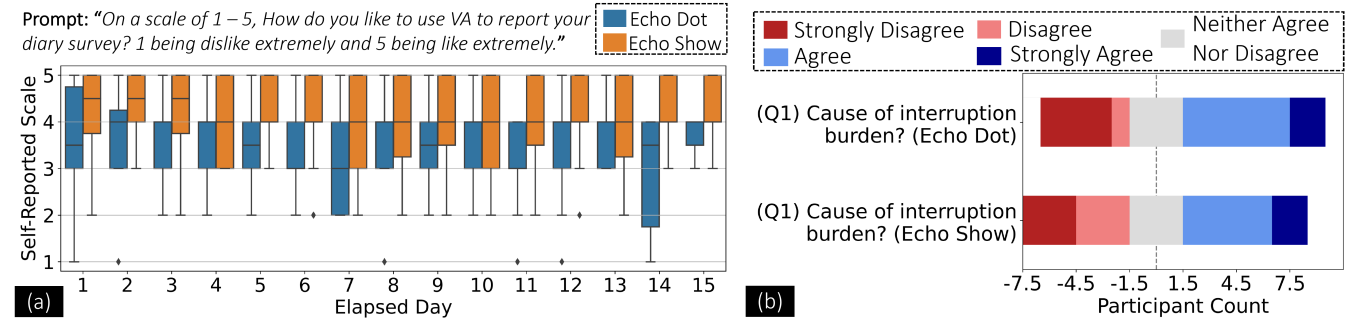


Figure 9: Evaluation results of (a) participants' responses of the usability prompt “on a scale of 1 – 5, how do you like to use the voice assistant to report your diary survey? 1 being dislike extremely and 5 being like extremely” reported through VAs in each study day, and (b) the 5-point Likert survey of how participant think the journaling of diary could cause interruption burden to the daily life routine (Q1). Incompleted responses were excluded from (a).

Echo Dot: “How many hours did you walk outside today?”

P8: “How many hours ... [unconsciously repeating the prompt causing the failures of valid response capturing]”

Additionally, due to the ambiguous nature of the speech conversation, participants suggested the need of designing additional ways to control the flow of the questionnaires (“I wanted something more interactive that we could go back and forth with kind of a conversation.” (P1)). Participants particularly wanted to have a way to go back and revise previous responses (e.g., “[it should] allow me to correct answers I’ve already said before” (P8)).

Visual output could be helpful for information consumption.

While the voice has been adopted as the major way for inputting information among older adults participants, many participants suggested the usefulness of having visual elements for information consumption, which might be one possible reason for Fig. 9a. One reason of such usefulness is that the question texts that are persistently shown on the touchscreen could help with older adults’ short working memory, despite that the designed diary questions could be easily responded fast. For example, “being old means my memory is shorter. By having the visual, it’s easier to keep moving along [while reporting the daily diaries]. Whereas with the Dot, you think you’re gonna say things, but then you forget the questions [...] With a visual. You have it [the diary questions] there and it keeps your mind focused on what you want to do or say” (P15). Similarly, P12 also mentioned

the usefulness of seeing possible Likert responses on the display: “I liked [the touchscreen], when the device asked a question and then it showed ‘one’, ‘two’, ‘three’, ‘four’ and ‘five’, and you could see. It’s easier for me to know and to remember. Whereas with the Dot, you have to remember what [the device] said, and sometimes [after the prompt being announced] I’ll think, oh.. what range did the Alexa tell me”.

While integrated reminders are needed for both devices, the larger form-factors of the Echo Show could lead to higher diary compliance. We encouraged participants to choose any methods they preferred for reminding themselves to conduct the daily diary survey. 13 participants reported that they remembered “just based on memory” or “memorized the task as part of their daily routines”, two participants used the reminder features on the standalone VAs or their smartwatch, one participant simply used his notes and calendar. While all participants initially felt confident about their selected reminders before the study, it turned out that participants still forgot. For example: “I thought it would be easy for me to remember to do it around meal times every evening because that’s when we get together in the kitchen. But unfortunately, I didn’t always remember to do that” (P10). Implicitly, P13 explained a possible reason of forgetting journaling daily diaries while the Echo Dot being covered by papers: “I just put my Alexa on the desk and I usually [completed the daily diaries] as long as I saw it. But there were couple of days missing when my desk was super messy and had my Echo Dot covered up by papers” (P13). This could be one possible

reason of the decreasing of compliance rate while using Echo Dot (Fig. 7), where forgetting is one major reason that cause failures of survey compliance.

4.3 RQ3: How Does the Built-In Touchscreen Affect the Older Adults' Behaviors and Experience of using VAs for General Purposes?

Our coding approach generated 31 codes categorized into four themes (Fig. 10a). Notably, while using the *Calling* service is out of our scope due to the needs of importing participants' contact books into our experimental accounts, which is not allowed by our IRB due to privacy concerns, we captured the intention of using such service by P13 and P16, leading to failure responses. Such findings were also verified by P13's questions during interviews of Phase 3: "can I make a phone call using Alexa? Can I say Alexa, call my sister Rose? [...] I can do that after the study, right?" Overall, we demonstrate a weak statistically increasing of measured interaction session while using the Echo Show compared to the Dot (ART: $F_{1,15} = 5.01$, $p = 0.041$, $\eta_p^2 = 0.25$, Fig. 10b). While no statistical significance were observed for interaction sessions related to *Chat*, increases of *Knowledge Query* (ART: $F_{1,15} = 6.88$, $p = 0.019$, $\eta_p^2 = 0.31$, Fig. 10c), *Service* (ART: $F_{1,15} = 9.82$, $p = 0.007$, $\eta_p^2 = 0.40$, Fig. 10d), and *Operation and Control* (ART: $F_{1,15} = 6.14$, $p = 0.026$, $\eta_p^2 = 0.29$, Fig. 10e) were observed. As for participants' self-reported survey results, we demonstrate a statistically higher overall SUS score ($F_{1,30} = 4.67$, $p = 0.040$, $\eta_p^2 = 0.13$, Fig. 11a) and a lower overall TLX score (ART: $F_{1,15} = 29.20$, $p < .001$, $\eta_p^2 = 0.66$, Fig. 11b) while using the Echo Show compared to Echo Dot, implying a better overall user experience. In particular, we measured the average overall SUS score yielded by using Echo Show being 75%, which is empirically considered as a good rating [63]. Fig. 11c shows that while most participants were comfortable and felt it was easy to use Echo Dot and Show, one participant held negative opinions. The qualitative analysis outlines our findings from three perspectives.

Participants preferred the lower disturbance level afforded by Echo Dot compared to Show. Participants reported how touchscreen caused visual disturbances during times of non-use: "[...] the screen changes all the time, and that can be irritating." (P2). While time displayed on screen was recognized to be useful, older adults emphasized that most visualizations could cause disturbances to some extent (e.g., "I would prefer to only show the time until I actually asked a question. But Amazon has prevented that from happening because they want to show you ads and other kind of things" (P2)). Further along this theme, two participants mentioned that the brightness of the display might affect the sleep quality when placed in the bedroom: "the Echo show is pretty bright. If you have it in your bedroom, I had to turn it toward the part of my desk [to avoid direct light]" (P6), "[the Echo Show] is on my nightstand, and I couldn't figure out how to control [...], so I got to turn that around to face another way" (P1).

Participants enjoyed seeing the additional auxiliary visual component on the display. Participants enjoyed the visual output together with the audio responses when using the device for general purposes. Examples include **auxiliary visual elements**

that are not announced by the voice output (e.g., "I like seeing the responses to questions. If you ask the Show, to add two numbers or multiply two numbers, it actually displays as well as telling you the answer [...] If you ask it about the weather. It'll tell you what the weather is going to be but it'll also show you little symbols [...]. So you get additional pieces of information from the screen that you don't get from the device without the screen" (P2)), **heterogeneous feature suggestions** (e.g., "the Echo Show is more preferable, because it's making suggestions about how it can be used" (P15)), **persistent visual information such as time and weather** ("the Echo Dot just sat there. Whereas the screen of the Echo Show gave me information about the day, the time, the weather, and also hints! It was like having a companion in the house that was silent, but still there! [...] I found it much more helpful and much more enjoyable as more than a simple device! [...] I enjoy getting up each morning and having it refresh me on the date, the time, and the weather, whereas for the Dot, I would have to ask them to give me the information. I guess I'm sort of lazy and prefer having it all out there for me" (P16)), and **explicit visual outputs for greetings and creating a sense of companionship** ("when I said 'thank you' to Alexa, there was a little blurb on the screen with the thank you message. That made me laugh and made me have a sense of companion" (P9)).

Many participants explained how visual outputs could make their generic uses of VAs become easier. First, most participants emphasized the usefulness of having visual components for specific types of information consumption. For example: "I love the display while listening to music because I can see the lyrics" (P5), "putting something [P9 added that the most medias experienced were Youtube videos] on screen to entertain me while cooking" (P9) and "when I wanted to see a recipe, the Echo Dot could not do that! It's a written thing that's laid out for me. But for Echo Show, it gives us a screen which can show a list of items. And that's one very useful way we could understand a recipe" (P10). Second, some participants believed the visual elements could help on reinforcing their memory regarding the output voice information. For example, "I didn't pay much attention to the screen during most of time, but I think it is always useful to have visual output to reinforce of what you're hearing" (P14). Finally, additional visual output might also compensate the hearing impairment though the visual sensory experience. For example: "if I am not close to an Alexa device, I don't always hear the words that she says. [With Show], the words will appear on the screen as well. So I can look at it, as well as hear it. [...] For people who have hearing aids and who don't necessarily wear them at home, Alexa can be very difficult to understand if you're not standing right next to her" (P2).

Impacts of failures of speech recognition. While participants enjoyed the conversational capabilities brought by voice + visual output, our results identified the impacts of ambiguity and recognition failures of voice commands. While using Echo Dot, P3 complaint: "occasionally, when I said 'Alexa! stop music', it would not stop the music. I had to unplug it and it was frustrated" (P3). Despite this, we found many participants subjectively think such occasional speech recognition failures will not cause significant impacts on the overall user experience. For example, "it happened about maybe 10% to 20% of the time [refer to the time when VAs cannot understand P13's intent] [...] It's usually because I talk too fast or my talking are not clear or something like that. It's not the devices fault [...] I think errors are mind and not about the machines" (P13).

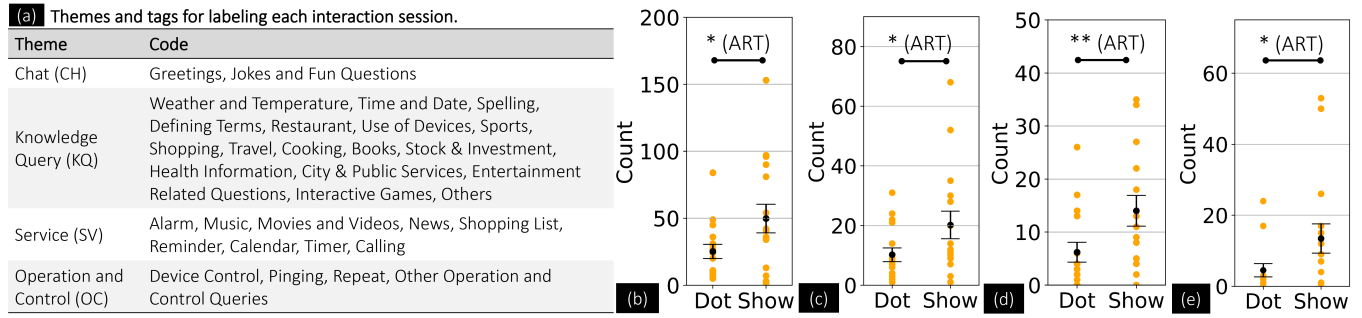


Figure 10: Characterizations of general uses. (a) Themes and codes used to label each interaction session. (b) The total number of interaction sessions being measured for each participant during the study for both devices. (c - e) The total number of interaction sessions related to Knowledge Query (KQ) (c), Service (SV) (d), and Operation and Control (OC) (e) while using both devices. Notations for indicating the statistical significance of *post-hoc* test: * = .05 > $p \geq .01$, ** = .01 > $p \geq .001$, *** = $p < .001$.

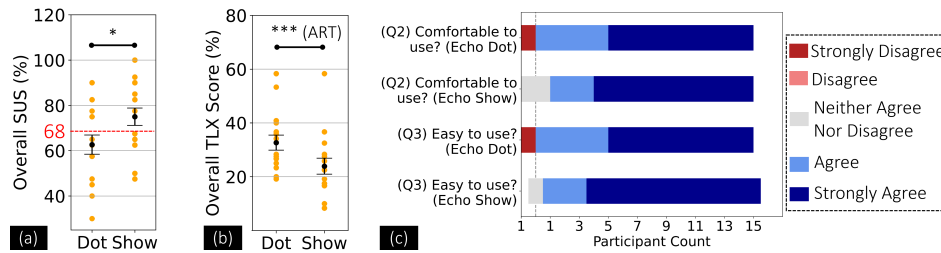


Figure 11: (a) Overall SUS [12, 13] and (b) overall NASA TLX [33] scores. A higher SUS and lower TLX score imply a better user experience. Notations for indicating the statistical significance of *post-hoc* test: * = .05 > $p \geq .01$, ** = .01 > $p \geq .001$, *** = $p < .001$. (c) The 5-point Likert survey results of how strong participants agree with that the device is (Q2) *comfortable to use* and (Q3) *easy to use*.

The location and placement of the display affect what and how to use it. While some participants conceptually acknowledged the usefulness of visual output and touch input, they also suggested how incorrect placement of the device could reduce its usefulness. During one focus group, P8 reasoned: “my device was stuck in a corner that was close to an outlet, and that was not particularly convenient [to access]”. P1 emphasized that being able to have hands-free interaction with the device was much more useful compared to the touch modality, possibly due to the non-optimal device placement, which could be one possible reason causing high physical and temporal task load in the post-study survey in Fig. 11. Example testimonies include: “it depends on where [the Show] is located. It was not easy to do the touch and it was much easier to use voice, because I can do it a little bit from a distance.” (P1), and “based on my experience so far, I would do it without touching. I’m lazy and it’s just easier! We had the device being about six feet away! You saw my office. It’s at the end of my desk and I sit over here in the corner, so it’s a lot easier to just yell over, such as ‘what time is it’ or ‘Alexa? Give me a 10 minute warning’ or something like that. So I would probably use it more if it was like right in front of me but my desk is so full now” (P13).

5 DISCUSSION

5.1 Using the Touchscreen for Device Setup and General Uses

Through our deployment study, we found that older adults held overall positive opinions for using the touchscreen as a secondary

modality on top of voice. With such insights, we offer three key design opportunities.

Integrating suggestions of device placement into the interactive device setup phases. The nature of hands-free and eyes-free interactions has pushed voice user interfaces as promising candidates for ambient assistive living technology, which eventually helps older adults better access and interact with inherently complex supporting technologies [29, 45]. While introducing a built-in touchscreen might enhance robustness and usefulness of the interaction system, the device accessibility might be degraded, as the interactions with touchscreen are not fully hands-free and eyes-free. Some older adults (e.g., P1, P8, and P11) pointed out that the placement of the devices could affect the general uses. Although being instructed to place the device in a preferred location, nearly all participants only considered the outlets’ location and the size of the device (e.g., “the cable is too short! I could only put it right here [points to an awkward place that is hard to reach]” (P7)). While setting up today’s voice + touchscreen VAs, there is no information regarding the potential degradation of the touchscreen-related experience caused by non-optimal placement of devices. Our finding implies that relying on only users’ intuition to place the voice + touchscreen VAs might not be feasible and could diminish the values of the touchscreen. Therefore, one future improvement could be to design ways to help users decide on device placements during the setup phase. Common features and suggested placement locations could also be crowd-sourced and provided to help older adults make the best decision.

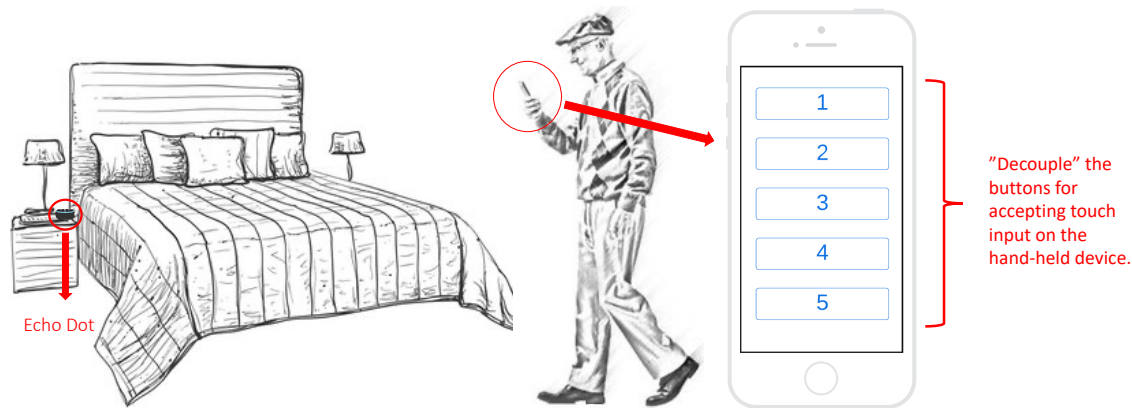


Figure 12: Conceptual design to use touchscreen for better control. Instead of using a built-in touchscreens, a multi-device system could combine a voice-only VA with a hand-held device used as an additional channel for inputting responses.

Maximizing the merits of visual components. Provided finding a suitable power outlets, we demonstrated a strong preference of participants to place their devices in an optimal location to maximize the accessibility of visual components (e.g., P2 put the Echo Show in front of her workstation, Fig. 1h, j). We also discussed how older adults appreciated the companion visualizations of the icons and texts of audio responses, during both diary journaling and generic uses. While Hu *et al.* [34] suggested that some older adults intended to bypass voice output and treat such voice-first VAs as touchscreen-only device (e.g., tablet) when it comes to incorrect speech recognition [77], we showed that older adults could, and are willing to, consume the “voice-first output” for general uses. While Han *et al.* [31] showed the promising of using touchscreen to visualize EMA data through a preliminary interview study, it is still unclear how to instantantiate such design tenet. Along this findings, future design might measure and consider *how*, *what*, and *when* to visualize, in order to maximize efficiency while consuming such “voice-first output”.

Design opportunities for context-awareness. While we showed how the touchscreen increases efficiency in terms of information consumption, some older adults emphasized the setbacks of disturbance, caused by, for example, the brightness of the screen at night and displaying irrelevant content at the focus time. Addressing such problems requires VAs that have the ability to adapt visualizations to real-world contexts. Example designs include dimming (or turning off) the display when older adults are sleeping or not in the room, and keeping the home screen visualizations consistent during focusing hours (e.g., only show the virtual clock). One future opportunity is to investigate how to detect and achieve context-awareness using less privacy-invasive sensors (e.g., light sensors, instead of cameras). This has also been previously suggested as one important concern by older adults who are adopting voice-first VAs [8, 18].

5.2 Using the Touchscreen to Keep Health Diaries

We showed that touchscreen affordances help older adults while journaling diaries, especially in terms of reducing response latency and increasing survey compliance. These findings lead to three key design implications.

Leveraging the affordances of touch input by decoupling primary and secondary input modalities. We showed that touch input leads to ~ 20% reduction in response latency, yet most participants still preferred to use speech input, mainly because the touch input was not always an accessible input modality. This implies that the built-in touchscreen supports older adults only for information consuming, instead of information input. Future design could focus on decoupling the input and output modalities. Fig. 12 illustrates an example where voice + touchscreen VAs can be extended into a multi-device system that spans across user-attached (e.g., hand-held device) and user-detached devices (e.g., voice-only VAs). While similar ideas have been proposed in TandemTrack [49] targeting general users who wanted to track exercise behavior, the differences between young and older adults in terms of technology proficiency and daily needs could pose unique challenges and needs to be investigated. For example, creating such a system requires designers to consider the implicit increment of interaction complexities and the impacts of device maintenance tasks (e.g., charging and configuration), which might in turn lead to poor usability. One future direction is to investigate the design trade-offs between enhanced accessibility of touch input and how usability could be sacrificed by the increased complexities brought by user-attached hand-held devices.

Opportunities to use the touchscreen for better control. We showed that using voice input alone could cause ambiguity and incorrect speech recognition, in particular for diary responses, and using the touchscreen as a secondary input support is promising to address this issue. While Fig. 12 might address these challenges by supporting higher interactivity on a handled device, our observations during device setup indicate that most older adults found it confusing to set up VAs using a decoupled smartphone. Therefore, instead of using voice-only VAs + smartphone, a touchscreen might still be needed to ease such process, but does not necessarily need to be integrated in the VA. Fig. 12 illustrates a “decoupled” touch-device that can be used as an independent add-on to the VA. Finally, introducing unnecessary features might increase a negative attitude among older adults [52]. To ease this problem, while still providing added functionality, future VAs should integrate more interactive guidance for helping older adults setting up devices and troubleshooting unexpected exceptions during device setup,

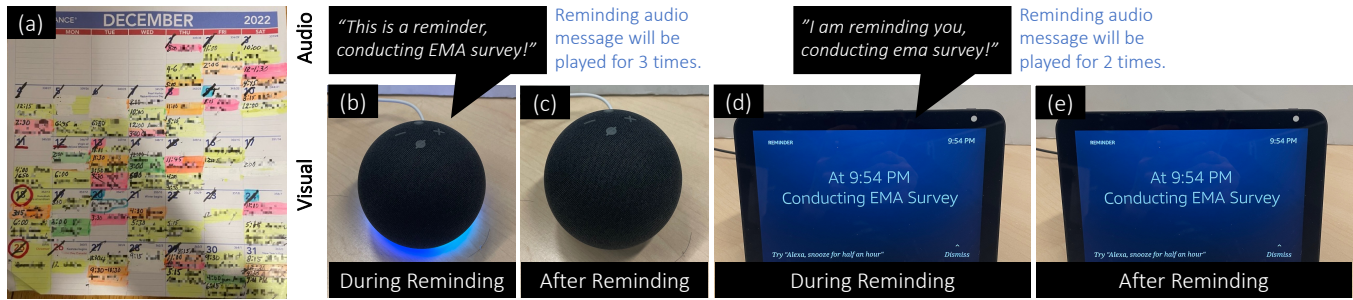


Figure 13: (a) An example paper-based note for event reminding from P12. Reminder features on standalone Echo Dot (b - c) and Echo Show (d - e).

and design metaphors for supporting more heterogeneous state navigation.

Opportunities to use the touchscreen for diary reminders and beyond. Part of our study focused on diary studies that do not have strict time requirement for participants to conduct the survey. Despite this, we asked participants to choose their preferred reminder methods for completing the survey. Most participants were confident that they would simply remember the tasks just based on their memory, as it would become part of their daily routine. However, we showed a different story: while measured survey compliance results (μ , 66.67% and 83.75% for Dot and Show) are comparable to the existing μ EMA systems where smartwatch vibration was used as the haptic notification [37] (81.21%), our participants reported that they started to forget. Designing effective reminding mechanisms for standalone devices is challenging due to the non-portability and the detached nature from end-users. Fig. 13b - c show how without touchscreen, the reminder message does not stay persistently after the initial triggering event. Even on Show, where visual reminder messages will stay persistently on the display, the proactive audio message will be snoozed after the event (see Fig. 13d - e). We, therefore, believe that leveraging a combination of standalone VAs with user-attached devices could be promising. In an ecosystem like the one shown in Fig. 12, proactive notifications could be designed on top of handheld devices without the need to set up a separate application. Besides diary journaling reminders, future work might also investigate how to bring older adults calendars (e.g., Fig. 13a) into such multi-device ecosystem.

6 LIMITATIONS

We recognize three limitations that might hinder the applicability of our findings to a more generalized setting. First, although our study was performed in a naturalistic environment, we evaluated only 16 older adults living in La Jolla in the United States. This may lead to biases in experiences that different populations might not have engaged with. Future work might investigate different groups of older adults who have a more varied experience with technology, live in different neighbourhoods, or share the VAs with their family members. Second, although we only focus on standalone VAs, their uses introduced inherent setbacks when users are more mobile (e.g., travel frequently). While assuming that older adults spend a considerable amount of time in their home, we did not assess interactions involving user-attached VAs when not at home. Future work could evaluate the ecosystem consisting of both user-attached

and user-detached devices and their interactions. Third, our current study used Echo Dot and Echo Show as the testbed due to the dominant market share (Sec. 3.3). While VAs from different vendors shared many similarities in terms of functions and designs, future work might consider evaluate other type of VAs in terms of older adults' perspectives. Finally, part of our research focuses on general uses of standalone VAs. However, under the restrictions of our current IRB protocol, we were not allowed to link our experimental account with the third-party services that need participants' private data (e.g., email and calling). Future deployment could consider older adults' private accounts (under different IRB protocols), which might offer more insights on their behaviors.

7 CONCLUSION

We conducted a within-subjects study ($N = 16$) using the Echo Dot and Show to understand how the voice + touchscreen VAs could influence older adults' experience of device setup, diary journaling, and general uses. Through a 40-day real-world deployment, we found that during the device setup, older adults appreciated the advantages of the touchscreen, with the overall TCT reduced by roughly 50% when using Echo Show compared to Echo Dot. As for diary journaling, while older adults enjoyed the visual output of touchscreen, they still preferred to respond to the prompts through speech, despite an approximately 20% of latency reduction while using touch input. Finally, we found that touchscreens were effective in encouraging older adults to engage more with VAs for general uses, despite the fact that input through touch was still referred to as not senior-friendly by our participants.

ACKNOWLEDGMENTS

This work is part of project VOLI and was supported by NIH/NIA under grant R56AG067393. Co-author Michael Hogarth has an equity interest in LifeLink Inc. and also serves on the company's Scientific Advisory Board. The terms of this arrangement have been reviewed and approved by the UC San Diego in accordance with its conflict of interest policies. We appreciate insightful feedback from the anonymous reviewers and discussions with colleagues from The Design Lab at UC San Diego, including Martin Yarmand, Janet G. Johnson and Manas Bedmutha. We thank Christopher Han and Peng Wei Lee for the help on the early stage implementations, and Mary Draper along with residents from the Vi at La Jolla for the help on participant recruitment.

REFERENCES

- [1] Samaneh Aminikhanghahi, Maureen Schmitter-Edgecombe, and Diane J. Cook. 2020. Context-Aware Delivery of Ecological Momentary Assessment. *IEEE Journal of Biomedical and Health Informatics* 24, 4 (2020), 1206–1214. <https://doi.org/10.1109/JBHI.2019.2937116>
- [2] Peter Anderberg, Shahryar Eivazzadeh, Johan Sanmartin Berglund, et al. 2019. A Novel Instrument for Measuring Older People's Attitudes Toward Technology (TechPH): Development and Validation. *Journal of Medical Internet Research* 21, 5 (2019), e13951.
- [3] Joan Palmiter Bajorek. 2018. *Voice First Versus the Multimodal User Interfaces of the Future*. <https://www.uxmatters.com/mt/archives/2018/10/voice-first-versus-the-multimodal-user-interfaces-of-the-future.php>
- [4] Tiago Carneiro Gorgulho Mendes Barros and Rodrigo Duarte Seabra. 2020. Usability Assessment of Google Assistant and Siri Virtual Assistants Focusing on Elderly Users. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*, Shahram Latifi (Ed.). Springer International Publishing, Cham, 653–657.
- [5] Timothy Bickmore and Justine Cassell. 1999. Small Talk and Conversational Storytelling in Embodied Conversational Interface Agents. In *AAAI Fall Symposium on Narrative Intelligence*. 87–92.
- [6] J Martin Bland and Douglas G Altman. 1995. Multiple significance tests: the Bonferroni method. *Bmj* 310, 6973 (1995), 170.
- [7] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003. Diary methods: Capturing Life as It Is Lived. *Annual review of psychology* 54, 1 (2003), 579–616.
- [8] Karen Bonilla and Aqueasha Martin-Hammond. 2020. Older Adults's Perceptions of Intelligent Voice Assistant Privacy, Transparency, and Online Privacy Guidelines. USENIX Association.
- [9] Walter R Boot, Neil Charness, Sara J Czaja, Joseph Sharit, Wendy A Rogers, Arthur D Fisk, Tracy Mitzner, Chin Chin Lee, and Sankaran Nair. 2015. Computer Proficiency Questionnaire: Assessing Low and High Computer Proficient Seniors. *The Gerontologist* 55, 3 (2015), 404–411.
- [10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [11] Sean Latrelle Bravo, Cedric Jose Herrera, Edward Carlo Valdez, Clint John Poliquit, Jennifer C Ureta, Jocelynn Cu, Judith J Azcarraga, and Joanna Pauline Rivera. 2020. CATE: An Embodied Conversational Agent for the Elderly. In *ICAART (2)*. 941–948.
- [12] John Brooke. 2013. SUS: a Retrospective. *Journal of Usability Studies* 8, 2 (2013), 29–40.
- [13] John Brooke et al. 1996. SUS—A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.
- [14] Miriam Cabrita, Richel Lousberg, Monique Tabak, Hermie J. Hermens, and Miriam M.R. Vollenbroek-Hutten. [n. d.]. An Exploratory Study on the Impact of Daily Activities on the Pleasure and Physical Activity of Older Adults. 14, 1 ([n. d.]), 1. <https://doi.org/10.1186/s11556-016-0170-2>
- [15] Kemeberley Charles, Chen Chen, Janet G. Johnson, Alice Lee, Ella T. Lifset, Michael Hogarth, Nadir Weibel, Emilia Farcas, and Alison A. Moore. 2021. How might an intelligent voice assistant address older adults' health-related needs?. In *Journal of the American Geriatrics Society*, Vol. 69. Wiley 111 River St, Hoboken 07030-5774, NJ, USA, S243–S244.
- [16] Chen Chen, Janet G. Johnson, Charles Kemeberley, Alice Lee, Ella T. Lifset, Michael Hogarth, Alison A. Moore, Emilia Farcas, and Nadir Weibel. 2021. Understanding Barriers and Design Opportunities to Improve Healthcare and QOL for Older Adults through Voice Assistants. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (Virtual Event, USA) (ASSETS '21)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3441852.3471218>
- [17] Chen Chen, Ella T. Lifset, Yichen Han, Arkajyoti Roy, Michael Hogarth, Alison A. Moore, Emilia Farcas, and Nadir Weibel. 2023. How do Older Adults Set Up Voice Assistants? Lessons Learned from a Deployment Experience for Older Adults to Set Up Standalone Voice Assistants. In *Designing Interactive Systems Conference (Pittsburgh, PA, USA) (DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 1–5. <https://doi.org/10.1145/3563703.3596640>
- [18] Chen Chen, Khalil Mrini, Kemeberley Charles, Ella Lifset, Michael Hogarth, Alison Moore, Nadir Weibel, and Emilia Farcas. 2021. Toward a Unified Metadata Schema for Ecological Momentary Assessment with Voice-First Virtual Assistants. In *CUI 2021 - 3rd Conference on Conversational User Interfaces (Bilbao (online), Spain) (CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 31, 6 pages. <https://doi.org/10.1145/3469595.3469626>
- [19] Y Choi, G Demiris, and H Thompson. 2018. Feasibility of Smart Speaker Use to Support Aging in Place. *Innovation in aging* 2, suppl_1 (2018), 560–560.
- [20] Dustin A. Coates and Max Amordeluso. 2019. *Voice applications for Alexa and Google assistant*. Manning Publications Co.
- [21] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.
- [22] Rebecca J Compton, Michael D Robinson, Scott Ode, Lorna C Quandt, Stephanie L Fineman, and Joshua Carp. 2008. Error-monitoring Ability Predicts Daily Stress Regulation. *Psychological Science* 19, 7 (2008), 702–708.
- [23] Delphine S Courvoisier, Michael Eid, Tanja Lischetzke, and Walter H Schreiber. 2010. Psychometric Properties of a Computerized Mobile Phone Method for Assessing Mood in Daily Life. *Emotion* 10, 1 (2010), 115.
- [24] Genevieve Fridlund Dunton, Keito Kawabata, Stephen Intille, Jennifer Wolch, and Mary Ann Pentz. 2012. Assessing the social and physical contexts of children's leisure-time physical activity: an ecological momentary assessment study. *American Journal of Health Promotion* 26, 3 (2012), 135–142.
- [25] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. *An Aligned Rank Transform Procedure for Multifactor Contrast Tests*. Association for Computing Machinery, New York, NY, USA, 754–768. <https://doi.org/10.1145/3472749.3474784>
- [26] Andrew Ennis, Joseph Rafferty, Jonathan Synnott, Ian Cleland, Chris Nugent, Andrea Selby, Sharon McIlroy, Ambre Berthelot, and Giovanni Masci. 2017. A Smart Cabinet and Voice Assistant to Support Independence in Older Adults. In *Ubiquitous Computing and Ambient Intelligence*, Sergio F. Ochoa, Pritpal Singh, and José Bravo (Eds.). Springer International Publishing, Cham, 466–472.
- [27] Centers for Disease Control and Prevention. 2021. *COVID-19 Risks and Vaccine Information for Older Adults*. <https://www.cdc.gov/aging/covid19/covid19-older-adults.html>
- [28] Tobias Goebel. 2020. *The Future Is Multimodal: Why Voice Alone Will Never Be the Answer*. <https://www.cmswire.com/digital-experience/the-future-is-multimodal-why-voice-alone-will-never-be-the-answer/>
- [29] Stefan Goetze, Niko Moritz, Jens-E Appell, Markus Meis, Christian Bartsch, and Jörg Bitzer. 2010. Acoustic user interfaces for ambient-assisted living technologies. *Informatics for Health and Social Care* 35, 3-4 (2010), 125–143. <https://doi.org/10.3109/17538157.2010.528655>
- [30] David H Gustafson, Marie-Louise Mares, Darcie C Johnston, Gina Landucci, Klaren Pe-Romashko, Olivia J Vjorn, Yaxin Hu, Adam Maus, Jane E Mahoney, and Bilge Mutlu. 2022. Using Smart Displays to Implement an eHealth System for Older Adults With Multiple Chronic Conditions: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols* 11, 5 (2022), e37522.
- [31] Yichen Han, Christopher Bo Han, Chen Chen, Peng Wei Lee, Michael Hogarth, Alison A. Moore, Nadir Weibel, and Emilia Farcas. 2022. Towards Visualization of Time-Series Ecological Momentary Assessment (EMA) Data on Standalone Voice-First Virtual Assistants. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (Athens, Greece) (ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 60, 4 pages. <https://doi.org/10.1145/3517428.3550398>
- [32] Margot J. Hanley and Shiri Azenkot. 2021. Understanding the Use of Voice Assistants by Older Adults. *CoRR* abs/2111.01210 (2021). arXiv:2111.01210 <https://arxiv.org/abs/2111.01210>
- [33] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [34] Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. 2022. Polite or Direct? Conversation Design of a Smart Display for Older Adults Based on Politeness Theory. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 307, 15 pages. <https://doi.org/10.1145/3491102.3517525>
- [35] Amazon Inc. 2020. *Echo Dot (3rd Gen) – Smart speaker with Alexa*. <https://www.amazon.com/Echo-Dot/dp/B07FZS74R>
- [36] Amazon Inc. 2020. *Echo Show 8 – HD smart display with Alexa – stay connected with video calling*. https://www.amazon.com/Echo-Show-Pantalla-inteligente-Alexa/dp/B07PF1Y28C/ref=sr_1_1?dchild=1&keywords=echo+show&qid=1598674780&sr=8-1
- [37] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. μ EMA: Microinteraction-Based Ecological Momentary Assessment (EMA) Using a Smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Heidelberg, Germany) (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1124–1128. <https://doi.org/10.1145/2971648.2971717>
- [38] Paul Jansons, Jackson Fyfe, Jack Dalla Via, Robin M Daly, Eugene Gvozdenko, and David Scott. 2022. Barriers and Enablers for Older Adults Participating in A Home-Based Pragmatic Exercise Program Delivered and Monitored by Amazon Alexa: A Qualitative Study. (2022).
- [39] Manuel Jesús-Azabal, José Agustín Medina-Rodríguez, Javier Durán-García, and Daniel García-Pérez. 2019. Remembranza Pills: Using Alexa to Remind the Daily Medicine Doses to Elderly. In *International Workshop on Gerontechnology*, José García-Alonso and César Fonseca (Eds.). Springer International Publishing, Cáceres, Spain, 151–159. https://doi.org/10.1007/978-3-030-41494-8_15
- [40] Yukta Karkera, Barsa Tandukar, Sowmya Chandra, and Aqueasha Martin-Hammond. 2023. Building Community Capacity: Exploring Voice Assistants to Support Older Adults in an Independent Living Community. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg,*

- Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 844, 17 pages. <https://doi.org/10.1145/3544548.3581561>
- [41] Sunyoung Kim. 2021. Exploring How Older Adults Use a Smart Speaker-Based Voice Assistant in Their First Interactions: Qualitative Study. *JMIR mHealth and uHealth* 9, 1 (2021), e20427. <https://doi.org/10.2196/20427>
 - [42] Sunyoung Kim. 2021. Exploring How Older Adults Use a Smart Speaker-Based Voice Assistant in Their First Interactions: Qualitative Study. *JMIR Mhealth Uhealth* 9, 1 (13 Jan 2021), e20427. <https://doi.org/10.2196/20427>
 - [43] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring Older Adults' Perception and Use of Smart Speaker-based Voice Assistants: A Longitudinal Study. *Computers in Human Behavior* 124 (2021), 106914. <https://doi.org/10.1016/j.chb.2021.106914>
 - [44] Bret Kinsella. 2020. *Amazon Smart Speaker Market Share Falls to 53% in 2019 with Google The Biggest Beneficiary Rising to 31%, Sonos Also Moves Up*. <https://voicebot.ai/2020/04/28/amazon-smart-speaker-market-share-falls-to-53-in-2019-with-google-the-biggest-beneficiary-rising-to-31-sonos-also-moves-up/>
 - [45] Thomas Kleinberger, Martin Becker, Eric Ras, Andreas Holzinger, and Paul Müller. 2007. Ambient intelligence in assisted living: enable elderly people to handle future interfaces. In *International conference on universal access in human-computer interaction*. Springer, 103–112.
 - [46] Jarosław Kowalski, Anna Jaskulska, Kinga Skorupska, Katarzyna Abramczuk, Cezary Biele, Wiesław Kopeć, and Krzysztof Marasek. 2019. Older Adults and Voice Interaction: A Pilot Study with Google Home. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, UK) (*CHI EA '19*). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3312973>
 - [47] Rock Leung, Charlotte Tang, Shathel Haddad, Joanna McGrenere, Peter Graf, and Vilja Ingriany. 2012. How Older Adults Learn to Use Mobile Devices: Survey and Field Investigations. *ACM Trans. Access. Comput.* 4, 3, Article 11 (dec 2012), 33 pages. <https://doi.org/10.1145/2399193.2399195>
 - [48] Ella T. Lifset, Kemeberley Charles, Emilia Farcas, Nadir Weibel, Michael Hogarth, Chen Chen, J Johnson, and A Moore. 2020. Can an Intelligent Virtual Assistant (IVA) Meet Older Adult Health-Related Needs in the Context of a Geriatric 5Ms Framework?. In *Journal of the American Geriatrics Society*, Vol. 70. Wiley 111 River St, Hoboken 07030-5774, NJ, USA, S245–S246.
 - [49] Yuhuan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. *TandemTrack: Shaping Consistent Exercise Experience by Complementing a Mobile App with a Smart Speaker*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376616>
 - [50] Roger C. Mannell and Jiri Zuzanek. 1991. The Nature and Variability of Leisure Constraints in Daily Life: The Case of the Physically Active Leisure of Older Adults. *Leisure Sciences* 13, 4 (1991), 337–351. <https://doi.org/10.1080/01490409109513149> arXiv:<https://doi.org/10.1080/01490409109513149>
 - [51] Siddharth Mehrotra, Vivian Genaro Motti, Helena Frijns, Tugce Akkoc, Sena Büşra Yengeç, Oguz Calik, Marieke M. M. Peeters, and Mark A. Neerincx. 2016. Embodied Conversational Interfaces for the Elderly User. In *Proceedings of the 8th Indian Conference on Human Computer Interaction* (Mumbai, India) (*IHCI '16*). Association for Computing Machinery, New York, NY, USA, 90–95. <https://doi.org/10.1145/3014362.3014372>
 - [52] Tracy L. Mitzner, Julie B. Boron, Cara B. Fausset, Anne E. Adams, Neil Charness, Sara J. Czaja, Katinka Dijkstra, Arthur D. Fisk, Wendy A. Rogers, and Joseph Sharit. 2010. Older Adults Talk Technology: Technology Usage and Attitudes. *Computers in Human Behavior* 26, 6 (2010), 1710–1721. <https://doi.org/10.1016/j.chb.2010.06.020> Online Interactivity: Role of Technology in Behavior Change.
 - [53] Debbie S Moskowitz and Simon N Young. 2006. Ecological Momentary Assessment: What It Is and Why It Is a Method of the Future in Clinical Psychopharmacology. *Journal of Psychiatry and Neuroscience* 31, 1 (2006), 13.
 - [54] Phani Nallam, Siddhant Bhandari, Jamie Sanders, and Aqueasha Martin-Hammond. 2020. A Question of Access: Exploring the Perceived Benefits and Barriers of Intelligent Voice Assistants for Improving Access to Consumer Health Resources Among Low-income Older Adults. *Gerontology and Geriatric Medicine* 6 (2020), 2333721420985975.
 - [55] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
 - [56] Andraž Petrovčič, Walter R. Boot, Tomaž Burnik, and Vesna Vesna Dolničar. 2019. Improving the Measurement of Older Adults' Mobile Device Proficiency: Results and Implications from a Study of Older Adult Smartphone Users. *IEEE Access* 7 (2019), 150412–150422. <https://doi.org/10.1109/ACCESS.2019.2947765>
 - [57] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 214 (Nov. 2019), 21 pages. <https://doi.org/10.1145/3359316>
 - [58] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. *ACM Trans. Comput.-Hum. Interact.* 27, 4, Article 31 (Sept. 2020), 27 pages. <https://doi.org/10.1145/3373759>
 - [59] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3173574.3174033>
 - [60] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Trans. Comput.-Hum. Interact.* 29, 3, Article 25 (jan 2022), 41 pages. <https://doi.org/10.1145/3484221>
 - [61] Colin Robson. 2002. *Real world research: A resource for social scientists and practitioner-researchers*. Wiley-Blackwell.
 - [62] Nelson A Roque and Walter R Boot. 2018. A New Tool for Assessing Mobile Device Proficiency in Older Adults: The Mobile Device Proficiency Questionnaire. *Journal of Applied Gerontology* 37, 2 (2018), 131–156.
 - [63] Jeff Sauro. 2021. SUSified? Little-Known System Usability Scale Facts. <https://uxpamagazine.org/sustified/>
 - [64] Marcia Shade, Kyle Rector, Kevin Kupzyk, et al. 2021. Voice Assistant Reminders and the Latency of Scheduled Medication Use in Older Adults With Pain: Descriptive Feasibility Study. *JMIR Formative Research* 5, 9 (2021), e26361. <https://doi.org/10.2196/26361>
 - [65] Shradha Shalini, Trevor Levins, Erin L. Robinson, Kari Lane, Geunhye Park, and Marjorie Skubic. 2019. Development and Comparison of Customized Voice-Assistant Systems for Independent Living Older Adults. In *Human Aspects of IT for the Aged Population. Social Media, Games and Assistive Environments*, Jia Zhou and Gavriel Salvendy (Eds.). Springer International Publishing, Cham, 464–479. https://doi.org/10.1007/978-3-030-22015-0_36
 - [66] Samuel Sanford Shapiro and Martin B Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3/4 (1965), 591–611.
 - [67] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological Momentary Assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.
 - [68] Brodrick Stigall, Jenny Waycott, Steven Baker, and Kelly Caine. 2019. Older Adults' Perception and Use of Voice User Interfaces: A Preliminary Review of the Computing Literature. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction* (Fremantle, WA, Australia) (*OZCHI'19*). Association for Computing Machinery, New York, NY, USA, 423–427. <https://doi.org/10.1145/3369457.3369506>
 - [69] Arthur A Stone, Saul S Shiffman, and Marten W DeVries. 1999. Ecological momentary assessment. (1999).
 - [70] Google Store. 2020. *Google Nest Mini – Smart Speaker for Any Room*. https://store.google.com/us/product/google_nest_mini?hl=en-US
 - [71] Graham Thompson and Ashok Ganesan. U.S. Patent 7472162B2, Dec. 2008. Communication System Architecture for Voice First Collaboration.
 - [72] Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a Toy": Exploring Older Adults' Reasons for Using, Limiting, and Abandoning Echo. 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/10.1145/3313831.3376760>
 - [73] John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.
 - [74] Pooja Upadhyay, Sharon Heung, Shiri Azenkot, and Robin N. Brewer. 2023. Studying Exploration & Long-Term Use of Voice Assistants by Older Adults (*CHI '23*). Association for Computing Machinery, New York, NY, USA, Article 848, 11 pages. <https://doi.org/10.1145/3544548.3580925>
 - [75] Adrián Valera Román, Denis Pato Martínez, Álvaro Lozano Murciego, Diego M. Jiménez-Bravo, and Juan F. de Paz. 2021. Voice Assistant Application for Avoiding Sedentarism in Elderly People Based on IoT Technologies. *Electronics* 10, 8 (2021). <https://doi.org/10.3390/electronics10080980>
 - [76] Daniel Vogel and Patrick Baudisch. 2007. Shift: A Technique for Operating Pen-Based Interfaces Using Touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (*CHI '07*). Association for Computing Machinery, New York, NY, USA, 657–666. <https://doi.org/10.1145/1240624.1240727>
 - [77] Kathryn Whitenon. 2017. *Voice First: The Future of Interaction?* <https://www.nngroup.com/articles/voice-first>
 - [78] World Health Organization (WHO). 2021. *Ageing and Health*. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>
 - [79] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (*CHI '11*). Association for Computing Machinery, New York, NY, USA, 143–146. <https://doi.org/10.1145/1978942.1978963>

Appendix A DESIGN OF DIARY SURVEY

This section provides supplementary details of the design of diary questions in Sec. 3.2. While establishing the validity of diary survey questions and the design of clinically-relevant diary studies is *beyond* our scope and left for future work, all older adults participants were instructed to provide their responses attentively as they will be carefully studied. The goal was to ensure that the participants were actually spending efforts on carefully deciding the responses for each prompts, which eventually aimed to mock up a realistic real-world diary journaling experience. We have explored and discussed iteratively the diary survey with one geriatric domain expert. Their focus during the design phase was to create prompts that could easily be answered by older adults, and could offer interesting insights into older adults' daily life for healthcare providers. In addition to the eight themes centered around empirical guidance from the World Health Organization (WHO) [78], we also added five questions at the end of each survey to understand older adults' *in situ* user experience. Due to the anthropomorphic nature of conversational voice assistants, especially in terms of older adults users [57], we emphasized a detailed disclaimer in the informed consent to not rely on the device for medical advice and we used “*If this is an emergency, call 9-1-1!*” as a short welcome message for each daily survey to remind older adults to not try to use the experimental testbed as a tool for seeking emergency help. The full question set is displayed in Fig. 14.

Catalogue	ID	Occupation Before Retirement
Quality of Sleep	1*	On a scale of 1 to 5, how well did you sleep last night? 1 being terribly to 5 being great.
	2*	How would you describe your sleep quality of the previous night on a scale of 1-5? 1 being terrible to 5 being great.
	3*	On a scale of 1-5, how well-rested did you feel this morning? 1 being not at all to 5 being very well-rested.
Social Interactions	4*	On a scale of 1-5, how satisfied are you with the amount of social interaction today? 1 being very unsatisfied, to 5 being very satisfied.
	5*	Did you talk with your friends or family today?
	6*	Were you doing exercise just now?
Exercise	7*	How many hours did you exercise for today?
	8*	On a scale of 1-5, how satisfied are you with the amount of exercise you did today? 1 being very unsatisfied, to 5 being very satisfied.
Pain Management	9*	How would you assess your pain on a scale of 1-5? 1 being not at all, to 5 being the worst pain imaginable.
	10*	Did you consume alcohol today? [if “yes” intent is captured, launch Q11 or Q12]
Alcohol	11	What kind of beverages, for example, wine, beer, vodka?
	12	What number of alcoholic drinks did you have today?
Food	13*	How many servings of fruits or vegetables did you consume today?
	14*	How many sweets did you consume today?
Symptom	15*	Do you have any symptoms that bothered you today?
	16*	Did you skip any prescribed medications today? [if “yes” intent is captured, launch 19]
Medication Management	17*	Did you skip any non-prescription medications that you normally take regularly today? [if “yes” intent is captured, launch 19]
	18*	Did you take any medications that you don’t normally take today? [if “yes” intent is captured, launch 19]
	19	Why?
Usability	20	What activity were you doing prior to this conversation?
	21	On a scale of 1 – 5, How do you like to use voice assistant to report your diary survey? 1 being dislike extremely and 5 being like extremely.
	22	How would you improve the voice assistant to better meet your needs?
	23	Did you face any challenges while using voice assistants? [if “yes” intent is captured, launch 24]
	24	What challenges did you encounter?

Binary Response
 Likert Response
 Number Response
 Open Response

Figure 14: We designed the diary survey focusing on eight themes of older adults' general wellness based on suggestions from professional geriatricians and empirical guidance from WHO [78]. For each theme, we provided different paraphrased versions, and one question will be randomly chosen from those marked by “*”. We also included the usability section to understand participants' *in situ* user experience.

Appendix B GUIDING QUESTIONS FOR THE INTERVIEWS

This section provides supplementary details of the semi-structured interviews we conducted in Phase 2 after device setup, Phase 3 after each 15-day session with Echo Dot and Show, as well as the focus groups in Phase 4. Fig. 15 shows the guiding questions we used for semi-structured interviews or focus groups at different study stages. The responses and discussions to the guiding questions were expected to be open-ended and the participants were *not* expected to *only* answer the questions. Instead, we encouraged participants to expand their responses and tell us more about their experience, stories, and rationales.

Index	Guiding Questions
(Phase 2) Experience of Setting Up Devices	
Q1	Are are key challenges that you were facing while attempting to setup the Echo Dot (or Echo Show)?
Q2*	Are there any experience differences while attempting to set up both devices? And what are potential the benefits and frustrations that the additional displays could bring? <i>This prompt was brought up only after the participant setting up both devices.</i>
(Phase 3 & 4) Experience of Daily Diary Survey Journaling	
Q3	How comfortable and convenience do you feel while using the Echo Dot (or Echo Show) to report and journal daily diary? And in general, how do you feel to journal daily diary using voice (or voice + touchscreen) while comparing to traditional method (using web—based form, text message and calling your providers)?
Q4	What are the challenges and frustrations that you were facing while using the Echo Dot (or Echo Show) to journal daily diary over the past 15 days?
Q5	How do you remember to conduct diary survey and what are the main reasons that you forgot?
Q6*	While using the Echo Show to journal binary, number and Likert type questions, do you prefer to use touch input or voice input to submit your responses? <i>This prompt was brought up only after the participant completing the 15-day session using Echo Show and in the focus group in Phase 4.</i>
Q7*	Could you give us some brief comparisons & contrasts between two VAs, for reporting daily diary survey? <i>This prompt was brought up only after the participant completing the second 15-day session in Phase 3 and in the focus group in Phase 4.</i>
(Phase 3 & 4) Experience of General Features In Use	
Q8	In general, how comfortable do you feel while using the Echo Dot (or Echo Show) for general uses?
Q9	What are the challenges and frustrations that you were facing for general uses over the past 15 days?
Q10	What features have you used? And how do you like it?
Q11*	While using the Echo Show, do you like to interact with the device using touch input or voice input? <i>This prompt was brought up only after the participant completing the 15-day session using Echo Show and in the focus group in Phase 4.</i>
Q12*	Could you give us some brief compare & contrast between two VA(s), for general uses? <i>This prompt was brought up only after the participant completing the second 15-day session in Phase 3 and in the focus group in Phase 4.</i>

Figure 15: Guiding questions used for semi-structured interviews at different study stages. Notably, the prompts Q2, Q6, Q7, Q11 and Q12, annotated by *, were only discussed at specific moments in the study stage.

Appendix C CODEBOOK AND THEMES FROM INTERVIEW DATA ANALYSIS

Fig. 16 shows the codebook and themes yielded from qualitative analysis of interview data collected from Phase 2 of the study. Similarly, Fig. 17 shows the codebook and themes of qualitative analysis results from Phase 3 and Phase 4.

Theme/Code	Count (Phase 2)
Overall Experience	45
Preference of voice + touchscreen	12
Preference of voice only	3
Speed for setting up device(s)	3
Ease of setting up device(s)	7
Effectiveness and design of user manual	14
Needs of assistance	6
Typing Experience	11
Experience of typing on built-in touchscreen of the Echo Show	5
Experience of typing on the phone	3
More familiar with phone/tablet	3
Feedbacks	17
Guidance and streamlined experience	7
Visual feedback on built-in touchscreen (while setting up the Echo Show)	4
Visual feedback on phone/tablet (while setting up the Echo Dot)	1
Immediate feedbacks related to device status	3
Locating buttons while setting up Echo Dot	2
Form Factors	13
Small form-factors of the Echo Dot	4
Bulky form-factors of the Echo Show	3
Device placements and external factors	6

Figure 16: The codebook resulted from our qualitative analysis of study Phase 2, showing four themes (bold). The “Count” refers to the number of participants’ quote tagged with corresponding theme (or code). Notably, it is possible that more than one codes are assigned to a specific quote.

	Theme/Code	Count (Phase 3)	Count (Phase 4)		Theme/Code	Count (Phase 3)	Count (Phase 4)
Experience of Daily Diary Journaling	Touch and Voice Input	48	54	Experience of General Features In Use	Negative Affordances related to Touchscreen	10	9
	Using touch input could speed up prompt responding	18	8		Physical size	1	7
	Using visual output could help older adults with hearing impairment	-	3		Disturbance caused by brightness of the display	4	2
	Using voice input could be helpful for older adults with mobility impairment	6	19		Visual noise and distractions	5	-
	Using visual output could reinforce memory	6	4		Voice + Visual Interactions	53	45
	Factors related to device placements and locations	5	8		Overall preference of the visual	19	1
	No difference between Echo Dot and Echo Show	4	-		Benefits of visual elements that are not delivered through speech	7	10
	Other rationales regarding using touch and voice input	9	12		Helps reinforce memory	1	6
	Interactions and Control of Conversational Flow	22	8		Companionships	7	7
	General comments for the needs of more conversational	15	3		Comments on specific features and use cases	17	14
	Needs to expanding answer	1	-		Compared and contrast with phone and/or tablet	2	7
	Needs for correcting previous submitted responses	5	5		Voice Recognitions	7	-
	Timeout issues for allowing speaking full responses	1			Subjective feeling of the performance	6	-
	Memorizing and Reminders	20	20		Outcomes of recognition failures	1	-
	Methods of memorizing journaling daily diary	9	3		Placements & Form Factors	36	10
	Problems of forgetting journaling daily diary	7	5		Use only voice due low reachability	10	8
	Suggestions of needs of effective reminder	4	12		Factors related to the location of outlets and/or external factors	6	2
					Conspicuous of Echo Dot	15	-
					Method of deciding device placement	5	-

Figure 17: The codebook resulted from our qualitative analysis, showing three themes (bold) related to the topic of *Experience of Daily Diary Journaling*, and another three themes (bold) related to the topic of *Experience of General Features In Use*. The “Count” refers to the number of participants’ quote tagged with corresponding theme (or code). Notably, it is possible that more than one codes are assigned to a specific quote.

Appendix D ETHICAL DISCLAIMERS

This work has been approved by the Institutional Review Board (IRB). Before the study, all participants have been introduced and signed the informed consent as well as the video and audio recording consent. Upon completing the study, the devices were reset and awarded to the participants as incentive, which is around \$130 as the time of writing (May 3, 2023) for both Echo Dot and Show. During the co-design workshop (Phase 4), participants were allowed to disable their camera and/or rename their Zoom account as needed, if they were not comfortable with showing camera feeds and their name to other study participants. We have gained consents for academic publications of all figures that are involved with anonymous participants. All analysis data has been unlinked with Personal Identifiable Information (PII) as per regulated in our IRB, and were stored in a secure cloud storage service, which complies with the Health Insurance Portability and Accountability Act (HIPAA). Due to the impacts of COVID-19 during Phase 2 of our study that required in-person visit, all research assistants strictly obeyed the guidance and regulations issued by local health authorities (*e.g.*, wearing masks and having a negative COVID-19 PCR test before the visit).