Understanding User Preferences of Voice Assistant Answer Structures for Personal Health Data Queries

Bradley Rey
University of British Columbia
Kelowna, British Columbia, Canada

Yumiko Sakamoto
University of British Columbia
Kelowna, British Columbia, Canada

Jaisie Sin
University of British Columbia
Kelowna, British Columbia, Canada

Pourang Irani
University of British Columbia
Kelowna, British Columbia, Canada

ABSTRACT

Voice assistants (VAs) are becoming ubiquitous within daily life, residing in homes, personal smart-devices, vehicles, and many other technologies. Designed for seamless natural language interaction, VAs empower users to ask questions and execute tasks without relying on graphical or tactile interfaces. A promising avenue for VAs is to allow people to ask personal health data questions. However, this functionality is currently not widely available and answer preferences to such questions have not been studied. We implemented a pseudo-VA that handles personal health data questions, answering in three unique styles: minimal, keyword, and full sentence. In two online user studies, 82 unique participants interacted with our VA, asking varying personal health data questions and ranking answer structures given. Our results show a strong preference for full sentence responses throughout. We find that even though full sentence answers have the longest mean response time, they are still found to provide high quality and optimal behaviour, while also being comprehensible and efficient. Furthermore, participants reported that for personal health question and answering, VAs should provide technical and efficient interactions rather than being social.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; **Natural language interfaces**; **Sound-based input / output**.

KEYWORDS

voice assistant, voice user interface, personal health data, natural language

ACM Reference Format:
Bradley Rey, Yumiko Sakamoto, Jaisie Sin, and Pourang Irani. 2024. Understanding User Preferences of Voice Assistant Answer Structures for Personal Health Data Queries. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3640794.3665552

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665552

1 INTRODUCTION

The rapid integration of voice assistants (VAs) within an array of devices is transforming the way we interact with technology. VAs now offer seamless interaction across standalone smart-speakers, smartphones, and wearables. The current capabilities of VAs allow for the handling of general commands (e.g., controlling entertainment devices) and answering of basic questions (e.g., the current weather). Now, pushing the boundaries of what a VA is capable of, we pose a scenario: A person hiking may look to their smartwatch for their current pace. Unsure how their pace compares to past hikes, they ask the smartwatch VA "How does today's pace compare to my hikes in the last month?" This is a question that could offer a voiced response. Currently, however, this level of question and answering through VA interaction is not possible, yet is very plausible.

As on-board mobile and standalone device computation advances, along with the collection of broader personal data, VAs will evolve beyond their current role in managing routine tasks and questions. People may soon find themselves interacting with VAs to obtain complex information, seek personalized recommendations, or as we explore in this work, and in the brief scenario above, to query their personal health data. VAs have the potential to support quick and easy querying of collected personal health data to offer richer and more personalized insight. Achieving this quick and easy exploration which provides greater insight, however, requires specific research focus [39].

Looking at past work, research has explored the use of VAs for addressing general health knowledge questions [1, 29] and for their use in healthcare [14]. Furthermore, research has explored VA answer structures, however only for common tasks and questions [20]. Ultimately, we do not have an understanding of how VA answers are perceived specifically for personal health data questions; uncertainty about how to provide answers is one such element that limits the potential of VAs to adequately respond to users' growing information needs. This missing knowledge can even factor into user satisfaction and adoption of these systems all together [42, 43].

To address this gap we implemented a browser-based pseudo-VA, similar to prior work [20], which allowed for study participants to ask and receive answers to personal health data questions. Across two online studies, 82 participants interacted with the VA, asking questions and receiving answers to a combined total of 30 unique question-answer pairs. Within our studies, we explored three answer structures (i.e., Minimal, Keyword, and Full Sentence), each paired with questions from four personal health question response

types (i.e., Open, Range, Binary, Value) and six known personal health insight categories (i.e., Contextual, Preemptive and Proactive, Goal and Performance, Combination and Comparison, Historical and Trend, and Current Status). We found that in comparison to general VA question and answers, participants greatly preferred the Full Sentence answer structure for most all response types and insight categories. Moreover, Full Sentence answers allowed for clarity in the answer when data was ambiguous, yet remained efficient despite the longer mean response time compared with other answer structures.

Our contributions are two-fold: **C1:** Two studies, utilizing a custom-built browser-based pseudo-VA, allowing participants to explore personal health questions and answers across a total of 30 unique personal health question-answer pairs. **C2:** Empirical insights into perceived answer quality, behaviour, comprehensibility, efficiency, and preference for personal health data questions. From these findings, we discuss current implications and directions for future works.

2 RELATED WORK

2.1 Interaction with Voice Assistants

Voice assistants (VAs) are quickly being adopted for use, now with over half of Americans using a VA in their homes [32]. To provide reason for such adoption, current research on natural language interaction (e.g., speech) has highlighted its utility in facilitating micro and hands-free interaction [3], especially when the visual system is overloaded (e.g., while tracking a walk) [7] and within various simultaneous activities [34].

As we interact with a VA using natural language, research understandably explores the idea of personification and human-likeness in VA answers and dialogue [25], as well as social interactions [35]. However, in stark contrast to this, others have highlighted a need for technical systems [15], rather than human. Notably, increased human-likeness tends to increase trust and privacy concerns [8, 27], concerns that are at the forefront for personal health data [44, 47]. The contrast in human-like versus technical can greatly influence the experience of the VA interaction. As such, it is important to continue to uncover preferences and user experiences as new VA interactions, such as for personal health data queries, become possible.

2.2 Voice Assistants in Health Contexts

Conversational user interfaces (CUIs) encompass a variety of interactive systems designed to facilitate natural language and conversational interactions. CUIs are an emerging means for people to gain general health-related information [1, 6, 29, 46], to self-report health and fitness data [28, 36], and to fill out health-related forms when health literacy is low [22]. CUIs in health contexts can take various forms, such as text-based chatbots, virtual assistants, and voice-activated platforms [14, 23]. Each of these forms offer unique capabilities for supporting health-related interactions and at times offer solutions to potential hurdles such as mispronunciation and recognition of medical terms and credibility [5, 33].

VAs, as a subset of CUIs, are becoming increasingly adopted as they are an embedded technology in many of our smart devices. VAs in health contexts currently allow users to query general health topics and symptoms [6], often providing links to online sources. As personal health data monitoring and exploration becomes more commonplace [11], the intersection of VA interaction and personal health data querying will quickly become a reality. Despite this potential, current VA systems do not fully leverage their capabilities for question-answering tasks. Furthermore, we largely do not know how a VA should answer personal health data queries, or how a VA is viewed for such tasks.

2.3 Voice Assistant Answer Structures

Investigations into VA communication styles have often focused from user to device. However, the feedback and response styles of the VAs not only play into a major design principle [30], but also a critical role in users' perceptions and adoption of these devices [42]. Yet, the effect of machine-to-human expression has been under investigated. For example, while factors like interruption [17] and conversational repair [12] have been explored, to our knowledge, only one work has explored answered structures for general VA use [20].

By exploring marketed VA answer practices as well as conducting a user study to compare various answer structures, Haas et al. [20] provide a comprehensive analysis of the experiences and user preferences regarding different VA answers. They first found that commercial VAs opt to convey more humanistic and full sentence answers for many common questions and commands. Only for home automation, where the outcome of the VA interaction is also noticeable within ones environment, were shorter keyword answers used. Then, in their own user study, Haas et al. found that minimal answers (those that provide the answer and no other supplemental or contextual information) were preferred for most command-based interactions while keyword answers (those that provide the answer and a brief confirmation of the keyword in the request) were preferred for most query-based interactions. This highlights a more utilitarian use of a VA than is currently recognized, where brief and basic answers suffice. Yet, many of the general use questions and commands explored are not *personal*. More specifically, there is a need for research to explore and understand people's expectations regarding answer structures when querying personal health data through VAs. One limiting factor hindering the ability of VAs to meet users' increasing information demands is the uncertainty surrounding how to effectively structure various answer types for personal health data queries, which this paper explores.

3 USER STUDIES METHODOLOGY

We conducted two online user studies. Within our studies, we used the same apparatus, procedure and data collection described below for both, and derived from previous work [20]. We follow the same procedure as previous work so that we can make direct comparisons and reflections between our findings, focused on personal health data question and answers, and previously published findings, which focuses on currently afforded question and answers as well as task-based commands.

3.1 Apparatus

We built a browser-based pseudo-VA using Javascript and the WebSpeech API[1]. The WebSpeech API enables two important elements: (1) Speech recognition (i.e., recognizing a personal health data question) and (2) Speech synthesis, where speech (i.e., the answer to the question) can be vocalized, for which we used the "Google US English Female" voice. We chose the female voice for its similarity to the default voices current VAs use. The pseudo-VA used within the user studies can be demoed in Google Chrome at https://vaphdqa.github.io/vaphdqa/.

We utilize the term *pseudo* to describe our VA, as the functionality of our VA was limited to only handle questions desired within our user studies. As such, the VA was not fully functional as we would expect a commercial VA to be (e.g., Google Assistant, Siri, Amazon Alexa, etc.). While this limits capability, it provides experimental control. During interaction with our pseudo-VA, the participant's question would be recognized and then processed by checking for keywords (and varying synonyms) specific to each question. Only when our pseudo-VA recognized all required keywords, was the appropriate answer vocalized. If recognition could not be made (e.g., required keywords were missing) a response would encourage the participant to follow the prompt given and to try again.

When interacting with our pseudo-VA, a button on-screen was used to trigger the start of recognition, rather than utilizing keyword detection as many commercial VAs do. This design choice was done for privacy reasons. As such, participants had control of when recognition would begin, with recognition ending once the WebSpeech API recognized a natural stop in the question spoken. Furthermore, the text in the button would change to inform the participant when the VA was listening.

3.2 Procedure

Each study comprised three distinct stages: an introduction, the main trials, and demographic surveys. The procedure, VA recognition, and data collection were refined across two separate pilot phases for each study. In the first pilot phase, a single participant took part in the study while sharing their screen on Zoom and providing think-aloud feedback throughout. The second pilot phase involved three participants, none of whom took part in the first phase, who mirrored the procedures that our main study participants would follow. This process allowed us to refine the methodology and address any potential issues.

Within each of the studies, we include two attention check questions. One question was built into the pseudo-VA, mimicking a typical study trial. Specifically, all participants had to ask about their calorie intake for the day, to which the pseudo-VA responded with instructions that needed to be followed on the next screen. The second attention check question was simply slotted within a demographic survey, asking participants to choose a specific response.

Throughout the studies, the pseudo-VA was embedded into a larger Qualtrics survey and opened in Google Chrome. Once open, participants proceeded through the survey, as described below. The study procedure was approved by our institution's ethics review board. All participants provided informed consent prior to starting a study.

3.2.1 Study Introduction. Before starting the main trials, participants were guided through a study introduction process, designed to optimize interaction and comfort with the browser-based pseudo-VA and the study in general. The study introduction comprised of the following elements: (1) Voice input and output verification. Participants could test their input and output by continually interacting with the pseudo-VA, using preset and generic questions. (2) Study task. We provided a brief overview of the study's context and tasks. (3) Practice trials. Participants experienced three trials, during which they were prompted to check their calendar for the following day. Note, as with the main trials, we did not have access to a participant's personal data; throughout, generic data was used when providing responses. (4) Personal health data preference. To further enhance interest and focus within our study, we asked participants to choose between heart rate and step count data as their preferred question and answer topic for the remainder of the study. These topics for personal health data exploration were chosen for their popularity [2, 37], and can be seen as topics that are of interest throughout many daily contexts [38].

3.2.2 Main Trials. The main trials in our study were structured into blocks, with each block focusing on questions requiring a specific response type (i.e., Open, Range, Binary, and Value) or pertaining to a specific category of personal health data (i.e., Contextual, Preemptive and Proactive, Goal and Performance, Combination and Comparison, Historical and Trend, and Current Status) in each of Study 1 (see Table 1) and Study 2 (see Table 2) respectively.

To ensure balance and minimize potential order effect, both the question blocks and answer structures within a single question block were ordered using a Latin Square design. Participants were randomly assigned to a question block ordering and for each question block randomly assigned an answer structure ordering. Throughout, we ensured as best we could that an equal number of each order for both the questions and answers was shown across all participants in the study. This study design allowed for a systematic exploration of how different answer structures performed across various dimensions of personal health data queries.

Trials were grouped in threes, corresponding to each question block. Participants asked the same question for each trial in a question block while receiving each of the three differently structured answers. We used a repeated measures design, in which each participant had to ask each of the four or six different questions for each of the three answer structures. This results in a total of 12 or 18 question/answer interactions (trials) for each participant, for Study 1 and Study 2 respectively.

During each trial, participants were prompted to ask a personal health data question using their own words. After providing the question, participants heard an answer given by the VA. Following a successful interaction with the VA for each trial, participants were automatically directed to complete the User Experience Questionnaire Plus (UEQ+) [21, 41]. The UEQ+ survey provides insights into participants' subjective experiences and satisfaction levels with the voice assistant answer structures.

Following the work which created the UEQ+ survey for voice assistants [21], and previous work [20], we paired two semantic

[1]https://wicg.github.io/speech-api/

scales to compose a single experience quality factor. We chose two semantic differentials with the highest found loading from three scales: 1) Behaviour, consisting of the scales *artificial - natural* and *unlikable - likeable*, 2) Comprehensibility consisting of the scales *complicated - simple* and *unambiguous - ambiguous*, and 3) Efficiency consisting *slow - fast* and *inefficient - efficient*. Finally, we incorporated a fourth scale: 4) Quality, which consisted of the scales with the third and fourth highest loading, yet still recommended by the creators [21]. The Quality scale consisted of the scales *unintelligent - intelligent* and *inappropriate - suitable*. We opted for these scales as the scales with the highest loading for Quality were *not helpful - helpful* and *useless - useful*. Through preliminary discussion, we felt these scales required the use of actual personal health data for a participant to fully evaluate these semantics.

After completing a question block, participants were asked to rank-order their preference for the three answer structures heard during that block. This ranking task aimed to elicit participants' subjective preferences with a forced response akin to selecting a single option if given the choice on their own device.

It is important to again note that all answers provided during the trials were generic, and no personal health data from participants was used in generating VA answers. This approach ensured consistency and privacy in the study design while still allowing for a thorough exploration of participants' preferences with respect to VA answer structures for personal health data queries.

3.2.3 Surveys and Open Feedback. Our study incorporated surveys and open feedback to gather insights into participants' demographics, personality traits, and preferences regarding voice assistant usage. We gathered information about age, gender, background, VA usage, and personal health data collection practices. Additionally, participants completed the Ten-Item Personality Inventory (TIPI) [19] to assess personality traits and the Attitudes For Technology Interaction (ATI) [18] scale to assess their attitudes and comfort with technology. We also explored attitudes towards voice assistants for both general and personal health data use, separating these surveys to mitigate carry-over responses. Finally, participants were given the opportunity to provide open feedback to express their thoughts, suggestions, and/or concerns.

3.3 Question Answer Structures

The questions prompted for participants to ask within each study were derived from a public dataset of personal health data queries captured in-the-wild from experienced smartwatch users [38]. The questions within our studies were carefully selected to represent the elicitation of different response types in Study 1 and various desired categories of personal health insights in Study 2. Within each study section below, we further highlight the questions used and how they were derived.

For both studies, the questions were answered using three answer structures: Minimal, Keyword, and Full Sentence. These answer structures have been utilized in previous work [20], and mimic the *Full* and *Brief* options offered by Google's Assistant. Minimal answers solely contain the information required to answer the question. Keyword answers provide the answer and confirmation of the question asked. Full sentence answers provide full sentence responses emulating human-like sentence structure. Notably, Minimal and Full Sentence answer structures follow humanistic response behaviours, while Keyword does not. As we allowed for both step count and heart rate as question topics to promote interest and engagement within the study, we aimed to ensure as much consistency as possible between the questions and answers for either data source. This included ensuring as much commonality between answers as possible while also ensuring answers were of similar lengths; see Table 1 and Table 2 for all questions and answers used (for both step count and heart rate topics). No matter the question topic chosen by participants, each participant saw the same number of question and answer trials during their respective study (i.e., 12 in Study 1 and 18 in Study 2).

3.4 Participant Recruitment

Recruitment of participants was done through Prolific[2]. Prior to participation, potential participants completed an eligibility survey to ensure they met our inclusion criteria: participants were required to (1) Have their first language be English. Therefore, answer structures could be properly evaluated by native speakers; (2) Have used a VA before. As such, participants held experiences either good or bad with the use of VAs and their answers; (3) Currently collect and/or explore personal health data. By having experience with personal health data, participants could have defined expectations and preferences.

When partaking in the main study, we asked participants to place themselves in a room with as little distraction as possible. We further required the use of a desktop/laptop computer with Google Chrome installed and for participants to have working microphone and speaker/headphones. No matter their responses, all participants who took part were paid 0.25 GBP for completing the eligibility and 6.5 GBP for completing the study.

3.5 Data and Analysis

The data from each study is analyzed across four key areas: (1) UEQ+ scores for each question-answer pair, (2) rank order preferences, (3) attitudes towards VAs, and (4) open feedback.

Important differences are highlighted for each study below with means and the confidence interval boundaries listed in text and graphically presented within figures. Throughout, we opt to calculate confidence intervals, using bootstrapping with a population estimate, rather than relying on p-values. Graphically presenting confidence intervals allows us to systematically assess any effects at play while also gauging practical significance [13, 16]. Reporting on confidence intervals rather than p-values has become increasingly popular in HCI literature [4]. Confidence intervals offer greater understanding for a broader audience and do not suffer from the illusion of truth sometimes provided by a p-value [16]. Before calculating the means and confidence intervals all outliers for each pairwise analysis were removed as any data point outside three standard deviations.

As the creators of the UEQ+ scale do not test for inter-item reliability, we calculated the Cronbach's α for each pairwise comparison using .70 as a cutoff [31]. When the α did not reach the cutoff for any UEQ+ scale, we report on the combined UEQ+ scale, while further exploring its semantic differentials separately.

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

Table 1: Questions and answers used in Study 1. The forward slash denotes the separation between the choice of heart rate or step count topics, one of which was chosen by the participant for use throughout the study.

Response Type	Expected Question	Minimal	Keyword	Full Sentence
Open	In what workouts does my heart rate reach zone five? / In what workouts does my step count reach 2,000 steps?	Indoor running, outdoor cycling, and rowing / Indoor running, outdoor running, and hiking	Workouts reaching zone five, indoor running, outdoor cycling, and rowing / Workouts reaching 2,000 steps, indoor running, outdoor running, and hiking	In the past, indoor running, outdoor cycling, and rowing workouts have brought your heart rate into zone five / In the past, you have reached 2,000 steps during indoor running, outdoor running, and hiking workouts
Range	What is my average heart rate on weekdays compared to weekends? / What is my average step count on weekdays compared to weekends?	85 beats per minute compared to 76 beats per minute / 9,820 steps compared to 10,680 steps	Weekdays, 85 beats per minute. Weekends, 76 beats per minute / Weekdays, 9,820 steps. Weekends, 10,680 steps	Your average heart rate during the week is 85 beats per minute. While on the weekends, your average heart rate is 76 beats per minute / Your average step count during the week is 9,820 steps. While on the weekends, your average step count is 10,680 steps
Binary	Is my heart rate higher than normal? / Is my step count higher than normal?	Yes / Yes	Yes, heart rate higher than normal / Yes, step count higher than normal	Your current heart rate is higher than your normal heart rate / Your current step count is higher than your normal step count
Value	What was my average heart rate in the last hour? / What was my step count in the last hour?	71 beats per minute / 1,375 steps	Average heart rate last hour, 71 beats per minute / Step count last hour, 1,375 steps	In the last hour, your average heart rate was 71 beats per minute / In the last hour, your step count was 1,375 steps
	Mean and Standard Deviation of Response Times (seconds):	M=2.6, SD=1.3 / M=3.0, SD=1.6	M=4.2, SD=1.1 / M=4.8, SD=1.4	M=5.7, SD=2.0 / M=6.7, SD=2.4

4 STUDY 1 - RESPONSE TYPES

4.1 Questions

For Study 1, the prompted questions were categorized by the type of response the question would elicit. To determine possible responses, we analyzed the utilized dataset [38]. Specifically, one coder created an initial code book, then two coders (including the original coder) utilized this code book to separately code a 10% subset of the dataset, and then compared codes. With an initial accuracy of 100%, the remainder of the dataset was separately coded and finally all codes were compared until a single code was applied to every question. From this process, four response types were found: 1) **Open**, a response requiring a list of data/information of any size, 2) **Range**, comparative two-value responses; 3) **Binary**, a yes/no response; 4) **Value**, a single value response.

Table 1 shows all questions and answers used within the study. Answers took an average of 2.8, 4.5, and 6.3 seconds to convey to the participant for each of Minimal, Keyword, and Full Sentence answer structures respectively.

4.2 Participants

Thirty-four participants took part, with one participant failing the attention checks. Of the 33 participants whose data we used for analysis their ages ranged from 18 to 63 years old (M = 35.0, SD = 11.7; 24 Females, 9 Males). Furthermore, 19 participants self-identified as White, nine as Black/African, three as Hispanic, and two as Asian. Twenty-two (22) participants indicated they use Google's Assistant, 17 use Apple's Siri, 11 use Amazon's Alexa , six use Samsung's Bixby, and two use Microsoft's Cortana. As marketed VAs offer similar responses [20], we were not concerned with bias from using a specific VA. Fifteen (15) participants stated using a voice assistant more than once per day, two once per day, eight a few times a week, two once a week, and six less than once per week. On average, participants had been collecting personal health data for 56.5 months (SD = 32.4 months). On average, Study 1 took 24.9 minutes to complete (SD = 11.2 minutes). Eleven (11) participants chose heart rate while 22 participants chose step count as their data type to explore.

4.3 Results

4.3.1 Quality, Behaviour, Comprehensibility, Efficiency. Mean participant ratings with 95% confidence intervals for answer quality, behavior, comprehensibility, and efficiency are shown in Figure 1.

Quality. For answer Quality (a composite of *helpfulness* and *usefulness*), significant differences were observed within Range, Binary, and Value response types, but not for Open. Within the Range response type, participants rated the Full Sentence answer structure as having the significantly highest Quality (M = 6.36, CI [6.11, 6.61]) compared to both Keyword (M = 5.77, CI [5.48, 6.07]) and Minimal (M = 4.94, CI [4.56, 5.32]). Furthermore, Keyword answers were rated significantly higher than Minimal. For the Binary response type, the Full Sentence answer structure received higher ratings (M = 6.11, CI [5.84, 6.38]) compared to both Keyword (M = 4.95, CI [4.56, 5.35]) and Minimal (M = 4.38, CI [3.79, 4.97]). For the Value response type, participants rated the Keyword answer structure significantly lower (M = 5.74, CI [5.43, 6.06]) when compared to the Full Sentence answer structure (M = 6.33, CI [6.09, 6.57]). Only when the response type was Open, the answer structure did not influence the perceived Quality.

We compared the quality ratings of Keyword answer structures across the four response types. In the Binary response type, Keyword answers were rated significantly lower (M = 4.95, CI [4.56, 5.35]) when compared to other response types. Similarly, for the Minimal answer structure, Binary responses received the lowest rating (M = 4.38, CI [3.79, 4.97]) when compared to Open and Value response types. While not significant, Full Sentence answers consistently received higher ratings across all response types, suggesting a potentially higher perceived quality.

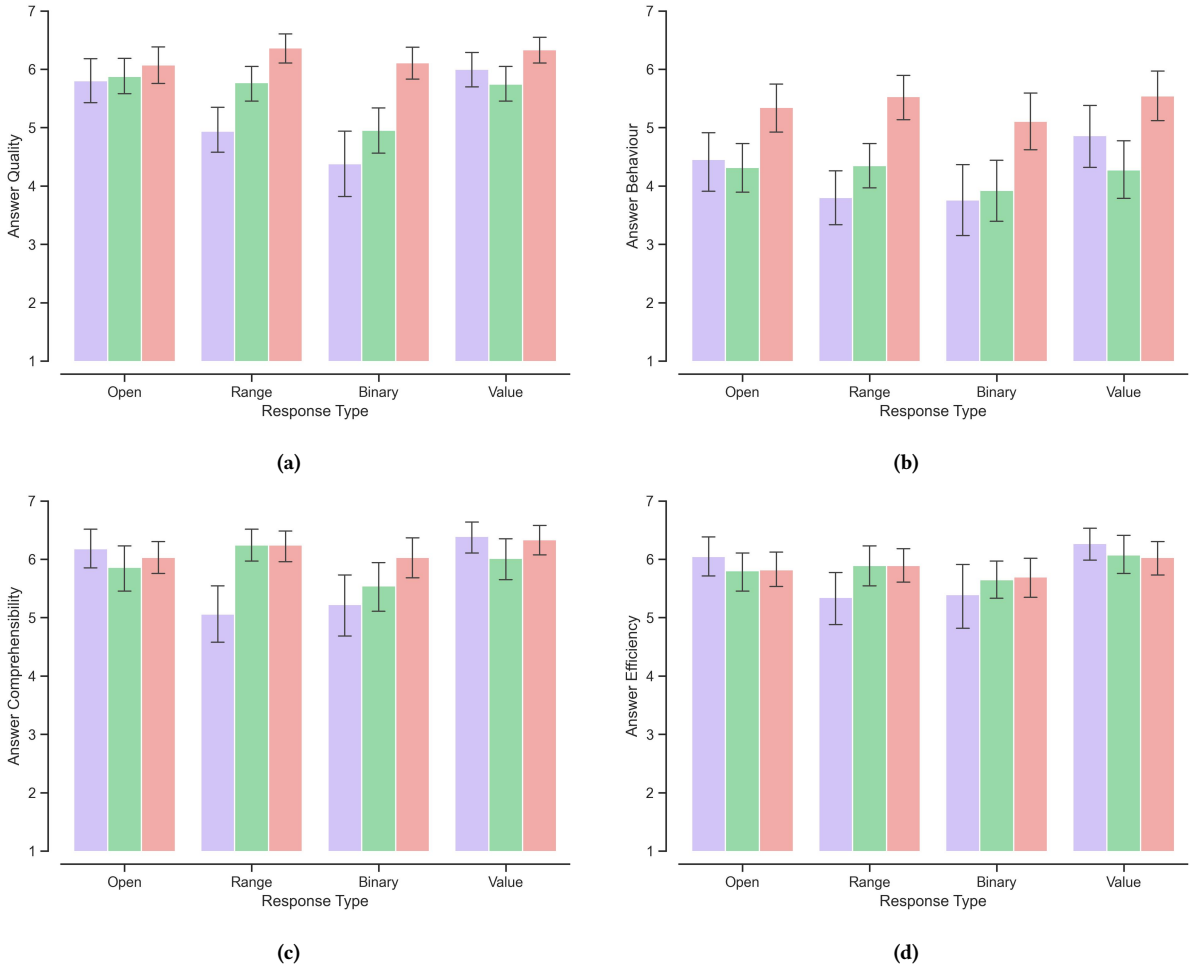


Figure 1: Study 1 mean UEQ+ ratings with 95% confidence intervals: Quality (a), Behaviour (b), Comprehensibility (c), and Efficiency (d). Ratings are compared by the response types (Open, Range, Binary, Value) and answer structures (within each group of three bars from left to right - Minimal, Keyword, Full Sentence) explored in the study.

Behaviour. In examining the answer Behaviour (a mean composite of *naturalness* and *likability*), a similar pattern for Full Sentence was evident. Regardless the response type, Full Sentence always yielded the highest ranking; significant differences between Full Sentence and the other answer structures within each response type were found for all except Value. Moreover, the means for Full Sentence did not vary across the four response types. This was consistent with both other two answer structure types (i.e., the Behaviour ratings were not influenced by Response Type). Thus, to further explore, we created mean scores for each answer structure type. As anticipated, the Full Sentence structure yielded the highest mean (M = 6.19, CI [5.99, 6.41]) while Minimum Sentence (M = 5.28, CI [4.97, 5.59]) and Keyword (M = 5.57, CI [5.32, 5.80]) did not vary.

Comprehensibility. In terms of answer Comprehensibility (a mean composite of *simplicity* and *ambiguity*), only one answer structure effect was found across response types. For the Range response type, Full Sentence and Keyword answer structures were seen as equally comprehensible while Minimal was seen as the least comprehensible (M = 5.06, CI [4.56, 5.56]). For other response types, the answer structure did not affect the level of comprehensibility. Noticeably, using a Minimal answer structure for both Range (M = 5.06, CI [4.56, 5.56]) and Binary (M = 5.23, CI [4.67, 5.79]) response types was seen as significantly less comprehensible than if used for Open and Value.

The level of Cronbach's α for Comprehensibility was 0.66. As such we separated the semantic differentials used and further explored *complicated - simple* and *ambiguous - unambiguous* separately; see Figure 2a. Significant differences across the semantic differentials are present for the Minimal answer structure in the Binary response type. More specifically, the Minimal answer was rated as being highly simple but is significantly different when compared to ambiguity, suggesting the response is simple yet ambiguous.

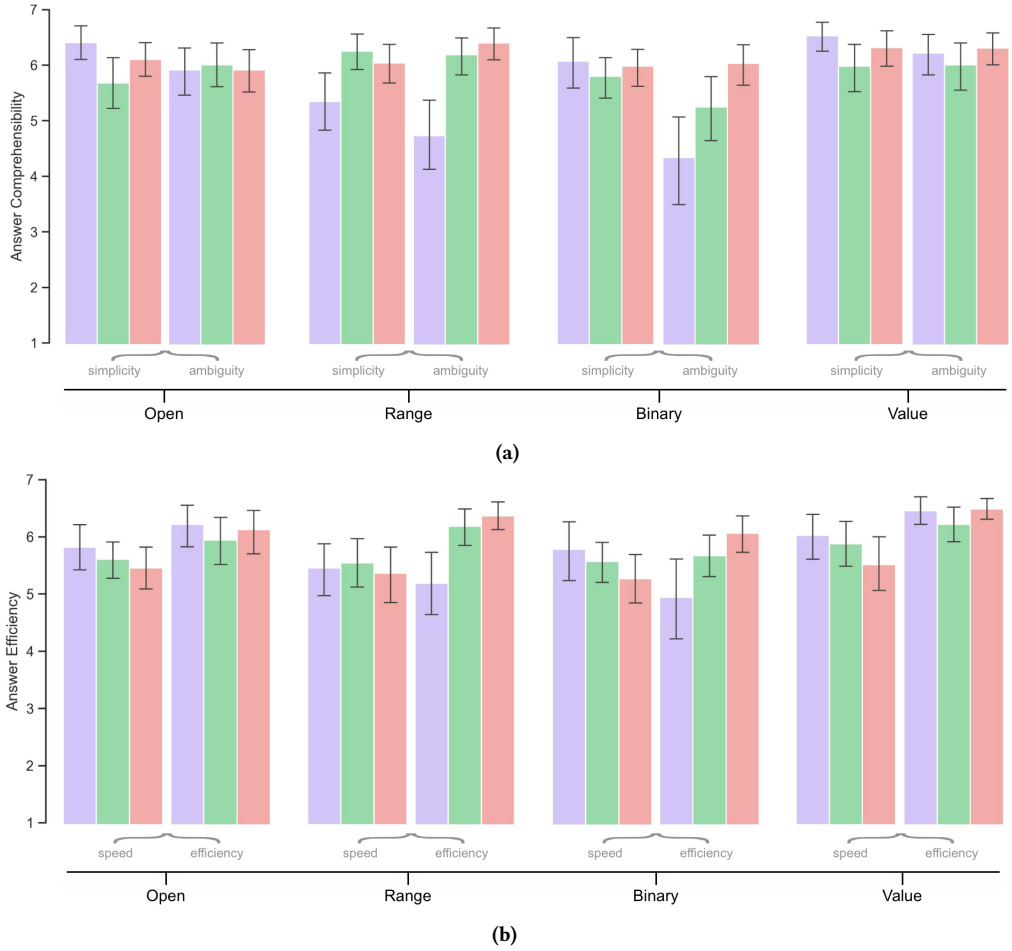


Figure 2: Study 1 mean UEQ+ ratings with 95% confidence intervals for a) *Comprehensibility* - separated by simplicity and ambiguity; and b) *Efficiency* - separated by speed and efficiency. Ratings are compared across the response types (Open, Range, Binary, Value) and answer structures (within each group of three bars from left to right - Minimal, Keyword, Full Sentence) explored in the study.

Efficiency. For answer Efficiency (a mean composite of *efficient* and *fast*), no differences were found.

However, the level of Cronbach's α for Efficiency was 0.65. As such we separated the semantic differentials used and further explored *slow - fast* and *inefficient - efficient* separately; see Figure 2b. Full Sentence answers were rated as significantly more efficient than they were fast for the Range, Binary, and Value response types. This suggests that while an answer does not have the fastest mean response times, as is the case with Full Sentence, they are still viewed as efficient by participants.

Interestingly, we realized that the sentence structure did not affect the perceived speed. VA's answers using Full Sentence, Keyword, and Minimal structures were perceived equally fast regardless of the actual response time (see Table 1 for mean response times).

4.3.2 Preference and Attitudes Towards Voice Assistants. Participants generally favored the Full Sentence answer structure; see Figure 3a. If Full Sentence answers were not preferred, then Minimal was often the preferred answer structure. This preference pattern suggests that participants prioritize responses that exhibit human structuring of answers. Notably, while participants demonstrated clear preferences for the Full Sentence answer structure throughout, only Range and Binary saw the majority of participants chose Full Sentence as their preferred answer structure. This is likely due to a need for additional context and clarity within these response types. In contrast, Value and Open response types allow for implicit interpretation and internal verification given an answer (i.e., if asking what workouts a person took 2000 steps within, cycling is an obvious wrong answer). This internal verification leads participants to perceive a slightly lesser need for Full Sentence responses; instead, favoring greater flexibility and brevity in the answer structure.

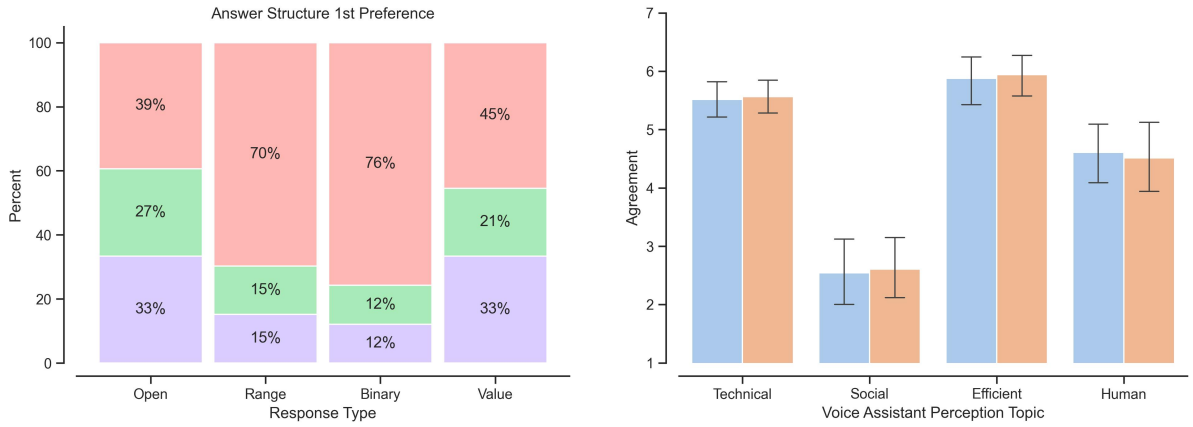


Figure 3: Participant's preference of answer structure (within each stacked bar from bottom to top - Minimal, Keyword, Full Sentence) for each response type (a). Participant's perceptions of voice assistants (within each group of two bars from left to right - VAs for general use and for personal health data queries (b).

Participants perceived voice assistants for personal health data exploration and general use similarly, indicating a consistent perception across these question domains; see Figure 3b. Interestingly, participants rated highly that a VA both was viewed as a technical system and should prioritize efficiency. Attributes related to human likeness and social interaction were not as strongly desired, with a VA social companion being even less preferred than human likeness. This suggests that participants value technical capabilities and efficiency in VAs, and may not necessarily expect or desire human-like or social qualities (e.g., as a trainer or coach).

4.3.3 Open Feedback. Across 18 unique comments, the open feedback provided by participants revealed three main topics. Firstly, two participants expressed a newfound interest in utilizing voice assistants for personal health data exploration. Second, regarding the study procedures, comments were generally positive, with seven participants expressing enjoyment and satisfaction. However, one participant suggested a need for a slower speed during interactions with the voice assistant. Additionally, two participants suggested the use of a different voice, specifically male, during interactions; no other interactive comments, such as recognition issues were mentioned. Finally, in terms of preferences, participants interestingly expressed diverse and opposing opinions. Two participants suggested favoring concise and clear responses, while another further mentioned they found human-like responses to be unsettling. Conversely, three participants preferred longer answers, particularly to confirm that the voice assistant understood their questions and attributing human-like qualities to these responses.

5 STUDY 2 - INSIGHT CATEGORIES

In our second study, we focus on insight categories of personal health data questions rather than broader response types. Notably, a question asked within a personal health data insight category can result in most response types, depending how the question is asked. Given the very few differences in UEQ+ scores between answer structures in both Value and Open responses in Study 1, we extend Study 1 by choosing to study Value responses in Study 2. This decision was driven by the versatility of Value responses, which are applicable across all insight categories, whereas Open responses are not. This study then offers a more nuanced understanding of user preferences in voice assistant answers for personal health data questions.

5.1 Questions

The prompted questions were chosen to represent known personal health insight categories [2, 9, 10, 24, 38]. These categories include: 1) **Contextual**, provides context to gain insight; 2) **Preemptive and Proactive**, provides insight into a future action; 3) **Goal and Performance**, derived from user goals and performance metrics; 4) **Combination and Comparison**, derived from combining/comparing data sources, time periods, and/or activities; 5) **Historical and Trend**, provides insight into past data; and 6) **Current Status**, derived from a current measured value. As the coded insight categories are not attached to the public dataset, we successfully reached out to Rey et al. [38] to ask for their coding as reported in their work. Subsequently, when choosing questions, we first opted for questions that were categorized into a single insight category. However, we note that some insight categories overlap with the Historical and Trend category.

Answers took an average of 1.8, 3.0, and 4.0 seconds for Minimal, Keyword, and Full Sentence answer structures respectively. As per our study goal, all answers invoked a Value response type. All questions and answers used within this study can be seen in Table 2.

5.2 Participants

Of the 52 participants who took part, three participants were removed for not passing our attention checks. Of the 49 participants whose data we used for analysis their ages ranged from 18 to 67 years old (M = 34.5, SD = 11.2; 30 Females, 19 Males). Furthermore,

Table 2: Questions and answers used in Study 2. The forward slash denotes the separation between the choice of heart rate or step count topics, one of which was chosen by the participant for use throughout the study.

Insight Category	Expected Question	Minimal	Keyword	Full Sentence
Current Status	What is my current heart rate? / What is my current step count?	68 beats per minute / 7,350 steps	Current heart rate, 68 beats per minute / Current step count, 7,350 steps	Your heart rate is currently at 68 beats per minute / Your step count is currently at 7,350 steps
Historical and Trend	What was my average daily heart rate last week? / What was my average daily step count last week?	78 beats per minute / 10,270 steps	Average daily heart rate, 78 beats per minute / Average daily step count, 10,270 steps	Your daily average heart rate last week was 78 beats per minute / Your daily average step count last week was 10,270 steps
Combination and Comparison	Is my heart rate different from my average? / Is my step count different from my average?	Higher than average / Higher than average	Current heart rate, higher than average / Current step count, higher than average	Your current heart rate is higher than your average heart rate / Your current step count is higher than your average step count
Goal and Performance	Which day of the week is my heart rate the highest? / Which day of the week is my step count the highest?	Saturdays / Saturdays	Highest heart rate, Saturdays / Most steps taken, Saturdays	Your heart rate is the highest on Saturdays / Your step count is the highest on Saturdays
Contextual	Is my heart rate lower in the morning, afternoon, or evening? / Is my step count lower in the morning, afternoon, or evening?	Evening / Afternoon	Lowest heart rate, evening / Lowest step count, afternoon	Your heart rate is the lowest in the evening / Your step count is the lowest in the afternoon
Preemptive and Proactive	How long should I control my breathing to get to my resting heart rate? / How far should I walk to get to 10,000 steps?	Two minutes / 2.5 kilometres	To reach your resting heart rate, two minutes / To reach 10,000 steps, 2.5 kilometres	To reach your resting heart rate, you should control your breathing for two minutes / To reach 10,000 steps, you should walk a distance of 2.5 kilometres
	Mean and Standard Deviation of Response Times (seconds):	M=1.6, SD=0.4 / M=1.9, SD=0.7	M=2.7, SD=0.4 / M=3.2, SD=0.7	M=3.6, SD=0.8 / M=4.3, SD=1.0

31 participants self-identified as White, 13 as Black/African, one as Hispanic, one as Asian, two as Multiracial, and one as Middle Eastern. Twenty-eight (28) participants indicated they use Google's Assistant, 23 use Apple's Siri, 26 use Amazon's Alexa, five use Samsung's Bixby, and four use Microsoft's Cortana. Twenty (20) participants stated using a voice assistant more than once per day, five once per day, 16 a few times a week, three once a week, and five less than once per week. Participants had been collecting personal health data for an average of 51.5 months (SD = 34.2 months). On average, Study 2 took 24.9 minutes to complete (SD = 9.4 minutes). Ten (10) participants chose heart rate while 39 chose step count as their data type to explore.

5.3 Results

5.3.1 Quality, Behaviour, Comprehensibility, Efficiency. Mean participant ratings with 95% confidence intervals for answer quality, behaviour, comprehensibility, and efficiency are shown in Figure 4.

Quality. Within insight categories, no significant differences in perceived quality are seen for Minimal and Keyword answer structures. However, Full Sentence shows a significantly higher mean Quality for Contextual (M = 6.15, CI [5.91, 6.38]) and Goal Performance (M = 6.31, CI [6.11, 6.5]) insight categories compared to other answer structures. As well, Full Sentence (M = 5.94, CI [5.7, 6.18]) shows a significantly higher quality than Minimal (M = 5.15, CI [4.81, 5.5]) in the Combination and Comparison insight category. Notably, across insight categories, there are no significant differences for each of the individual answer structures, suggesting that the insight category does not change perceived Quality of an answer structure.

Behaviour. In parallel to Study 1, we once again noticed a general trend wherein Full Sentence resulted in the highest scores for Behaviour (*naturalness* and *likability*). However, Minimal Sentence and Keyword, which generally scored lower than Full Sentence, did not vary from one another.

Comprehensibility. The level of Cronbach's α for Comprehensibility was 0.49. Thus, we investigated the semantics (*complicated - simple* and *ambiguous - unambiguous*) independently; see Figure 5. Noticeably, as in Study 1, it is the Minimal answer structure which provides significant differences comparing across the two semantics in the Contextual, Goal and Performance, Combination and Comparison, as well as Historical and Trend insight categories. Each time the answer is rated as significantly more simple while being ambiguous.

Efficiency. For Efficiency (a mean composite of "*efficient*" and "*fast*") no differences were found. The answer structure did not influence the levels of perceived efficiency. No other effects were found.

5.3.2 Preference and Attitudes Towards Voice Assistants. As with Study 1, we again see the Full Sentence answer structure as being the preferred majority for the remaining insight categories; see Figure 6a. Only for the Current Status insight category is this less pronounced. As complexity in the question increases, from that of a Current Status question, Full Sentence seems to be preferred mainly for its Quality and Behaviour. Finally, perceptions of VAs across both studies were comparable; see Figure 6b

5.3.3 Open Feedback. Across 19 unique comments the same three topics arose from Study 1. First, three participants expressed an

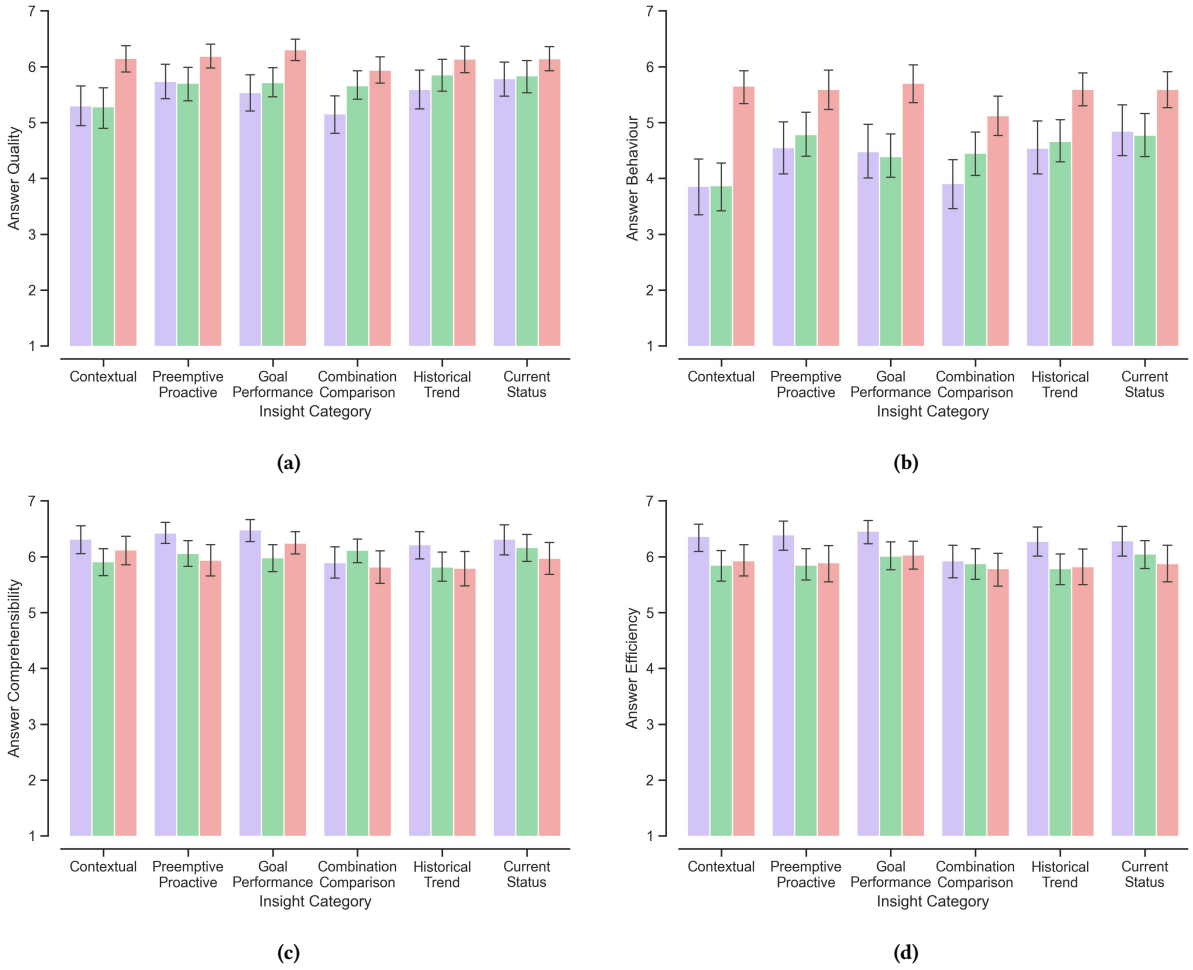


Figure 4: Study 2 mean UEQ+ ratings with 95% confidence intervals: Quality (a), Behaviour (b), Comprehensibility (c), and Efficiency (d). Ratings are compared by the insight categories (Contextual, Preemptive and Proactive, Goal and Performance, Combination and Comparison, Historical and Trend, and Current Status) and answer structures (within each group of three bars from left to right - Minimal, Keyword, Full Sentence) explored in the study.

interest in using a VA for personal health data exploration. However, two participants shared that they prefer to perform visual data analysis. These comments are important; our aim is not to replace visual data analysis. Instead, our work aims to diversify approaches to explore personal health data. Second, nine study procedure comments highlighted that the study went well. One participant mentioned the VA could slow down its answers while only three participants expressed the VA was *sensitive* which caused some interpretation issues throughout (sensitivity was expressed as being due to background noise, a learned accent, and an illness influencing speech). Finally, only one participant commented that they preferred concise and clear answers so as minimize the time taken for the interaction.

6 DISCUSSION

6.1 Implications for the Design of VA Interactions for Personal Health Data Queries

6.1.1 Comparisons with General VA Interactions. Our results indicated that users preferred Full Sentence answers for their personal health data queries. This runs counter to prior work, which has suggested that Minimal and Keyword responses are ranked positively, and sometimes preferred, for common VA tasks (i.e., knowledge queries, home automation, reminders, calendar queries) [20]. We believe this discrepancy arises due to the brevity of Minimal and Keyword responses which fall short in conveying the level of comprehension required for many personal health data queries. For example, if a person asks "What is on my calendar tomorrow?", the VA could respond using a Minimal answer structure, stating "Lunch

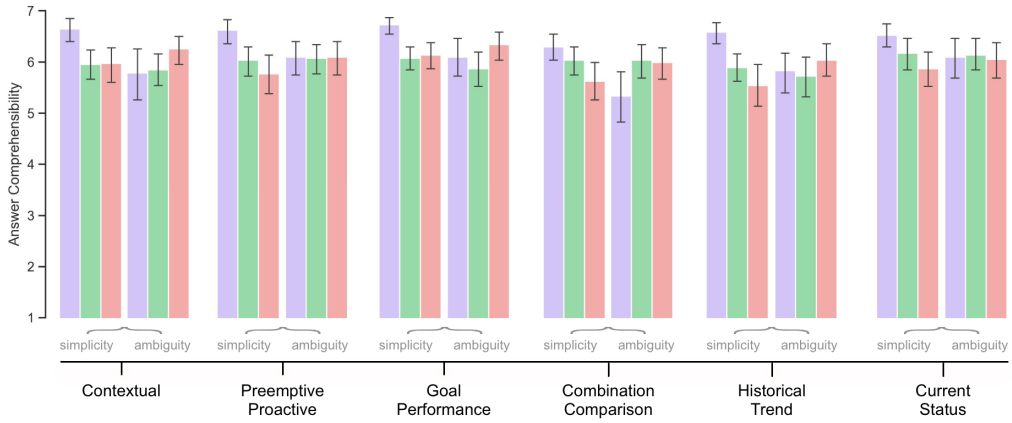
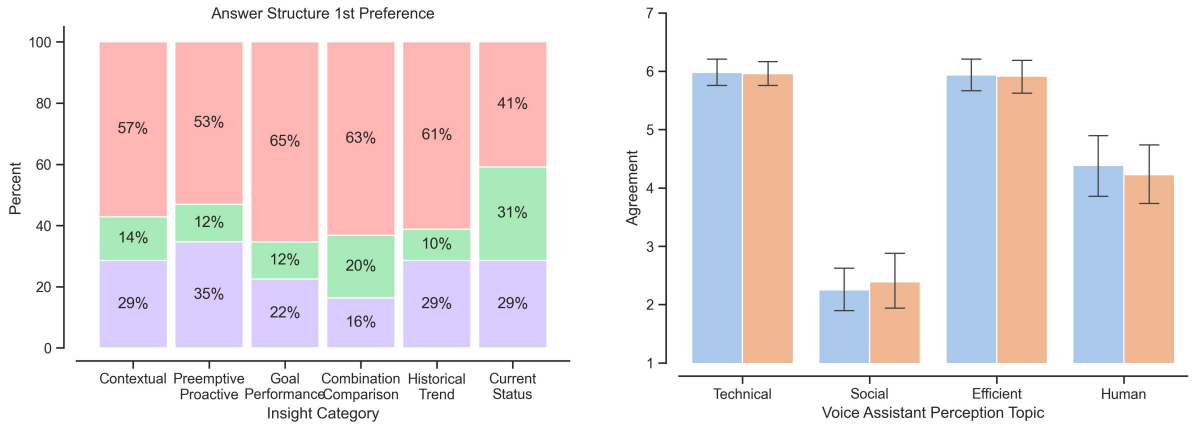


Figure 5: Study 2 mean UEQ+ ratings with 95% confidence intervals for Comprehensibility - separated by simplicity and ambiguity. Ratings are compared across the insight categories and answer structures (within each group of three bars from left to right - Minimal, Keyword, Full Sentence) explored in the study.



(a) **(b)**

Figure 6: Participant's preference of answer structure (within each stacked bar from bottom to top - Minimal, Keyword, Full Sentence) for each insight category (a). Participant's perceptions of voice assistants (within each group of two bars from left to right - VAs for general use and for personal health data queries (b).

with Tyler, 1PM. Games with Danica, 7PM." The content in the answer itself provides a connotation of calendar events and does not produce an answer that is ambiguous. In contrast, if a person asks "What is my average daily step count in the last week?" Providing a Minimal answer, such as "10,320 steps", does little to convey that the question was properly understood. Many other possibilities exist for a similar answer (e.g., average daily step count in the last month or current step count.) Such ambiguity has been previously noted as a barrier in using personal health data for clinical purposes [45], and now appears to be a common barrier for end-users exploring their own personal health data.

Our findings highlight the importance of tailoring answer structures to specific questions, both general and health-related. Furthermore, contextual information plays a key role in VA personal health data exploration, where users comprehend the information provided holistically, rather than focusing solely on single numerical or categorical values. As such, design considerations for VA personal health data interactions should prioritize confirmation and inclusion of key aspects of the data.

6.1.2 Efficient Full Sentence Answers. Despite being longer with respect to response time, Full Sentences were perceived as equally efficient as Minimal and Keyword answers. We contemplate several reasons for this observation. First, contextual information provided within a Full Sentence may contribute to a more comprehensive understanding of the answer, thereby reducing the need for follow-up questions or clarification. This can ultimately enhance efficiency, even if individual responses take longer. Second, the context in which the answer is given may influence its perceived efficiency. In distracting environments, concise responses may be preferred,

while in quieter settings, more detailed answers may be deemed appropriate. Thus, the threshold of an efficient answer may vary depending on the situational context, allowing for flexibility in response length without compromising perceived efficiency.

Not only do our findings indicate that Full Sentence is perceived to be efficient, but they suggest that there may be room to augment Full Sentence answer content without sacrificing perceived efficiency. This is due to the fact that we observed that participants *feel* Full Sentence answers were equally as fast as Keyword and Minimal answers. Looking ahead, leveraging the capabilities of Full Sentence answers could allow for serendipitous information, akin to visual data exploration. For example, if a person asks "What is my average daily step count in the last week?", the VA could provide a Full Sentence answer stating, "Your average daily step count in the last week is 10,320 steps. This is higher than the previous week, keep it up!" Moreover, future investigations could aim to enhance the depth of information conveyed, the number of data points included, and/or the influence of certain answers to enrich user interactions with VAs. For example, if a person asks "Is my heart rate lower in the morning, afternoon, or evening", the VA could respond with "Your heart rate is the lowest in the evening, and is roughly 15 beats per minute lower than other times of day." Each of these areas of exploration should explore the context in which VA interactions take place (i.e., at home or while on a walk). Future work in these areas can build from the results found in our work while then aiming to provide more comprehensive guidelines for personal health data questions and answers, allowing for VA interactions that suit people's needs and preferences and could provide greater influence.

6.2 Human Emulation and Unwavering Perceptions

Our results highlighted unwavering and confident perceptions within both of our studies for the use of VAs for personal health data question and answering. This can be seen in the highly similar responses captured in our VA perception survey questions (see Figure 3b and Figure 6b, and follow the results of previous work exploring general VA use [20]. As seen from this reported data, striking a balance between technicality and efficiency, while providing answers that emulate full sentences (and therefore human likeness) is key. Importantly, however, we must be cognizant that VA interactions should remain to invoke as little social interaction as possible (e.g., the VA should not emulate a fitness coach). Notably, VA responses which emulate human behaviour, in part conveyed through Minimal and preferred Full Sentence answer structures, have been shown to raise people's expectations of the VA [26]. While expectation and capability may be a concern in the earlier life cycles of VAs, rapid improvements to VA performance will likely mitigate these concerns over time. Of more interest, is that human-like responses can lead to incorrect and inappropriate use of a VA [40]. Furthermore, trust and privacy become concerns when the VA is seen as increasingly human-like [8, 27], a concern that is amplified with respect to personal data [44, 47] over that of general knowledge (e.g., the weather or population of the USA). Therefore, we encourage designers of VA technologies to pay close attention to the balance required to accommodate these preferences.

6.3 Comparisons With Commercial Personal Health Data Question and Answering

To better situate our work we explore current capabilities with respect to VA personal health data question and answering. To our knowledge, Apple's Siri is the only commercial VA which can answer some personal health data questions. Other commercial VAs often recognize key words and provide a prompt to open a respective health app. As such, we asked Siri a range of personal health data questions. As functionality is still limited (i.e., Siri can not answer many of the questions within our study), we explored questions, and slight variations of questions promoted by Apple[3]; see Table 3.

Notably, this is a new feature (as of December 2023) that only works with iPhones and iPads running iOS and iPadOS greater than 17.2 and Apple Watch Series 9 and Ultra 2 running watchOS greater than 10.2. The use of Siri for the exploration of personal health data is coupled with a display, rather than through a standalone device (e.g., in a standalone smart speaker and as seen in our study). As such, we provide this information for discussion purposes only.

Our study findings reveal preferred differences compared to how Siri answers questions. While Siri predominantly uses Minimal vocal responses and Keyword information displayed on screen, our studies highlight a preference for Full Sentence answers. Notably, Siri rarely employs Full Sentences, except for activity ring data, where multiple data points are conveyed, and when there are contextual deviations from the question (see Table 3, "What's my heart rate?" - heart rate data was not current and Siri responded with the last sample recorded). This divergence from our study findings in VA practice highlights varied strategies in voice assistant design, with Siri prioritizing brevity and visual support. However, the presence or use of a screen may not always be optimal (as in the example shared in the Introduction where the focus should remain on the hiking environment). Such insight and discussion sheds light on the diverse approaches that can be adopted by voice assistants in managing personal health data and more importantly underscores the importance of understanding user preferences, expectations, and contexts.

6.4 UEQ+ Semantic Differentials

In our study, despite using an adapted version of the the UEQ+ survey to measure VA user experience [21], we observed conflicting semantic differentials for Efficiency and Comprehensibility. The lack of correlation found suggests that the semantic differentials may not be effectively capturing the same intended user experience factor. While it can be argued that two items do not need to directly relate to provide a worthwhile assessment, as we do in this work, the divergent performance of these factors raises some concerns. For example, a response may be both simple and ambiguous, resulting in a lower overall comprehensibility. However, it is better to isolate factors that measure the same element of user experience for a more comprehensive and nuanced understanding. Our findings highlight the need for future work to refine the UEQ+ scale for VA interactions, aligning semantic differentials more closely with user experiences. Specifically, better assessing combinations of the

[3]https://www.apple.com/newsroom/2023/12/siri-can-now-help-users-access-and-log-their-health-app-data/

Table 3: Recorded personal health data question and answers using Siri on an iPhone 14 Pro running iOS 17.3.

Question	Vocal Response	Supplementary Information Displayed
What is my current step count?	[Step Count] steps	Steps Today
What's my heart rate?	As of February 2, 2024 11:48 AM, it was 77 BPM	Heart Rate [Month] [Day] [Year] [Time]
How far did I walk yesterday?	[Distance] km	Walking + Running Distance Yesterday
How far have I walked this week?	[Distance] km	[Distance] km Daily Average Walking + Running Distance [Month] [Day] - [Month] [Day] [Year]
What is my move ring at?	You've burned [Current Calorie Burn] out of your [Calorie Burn Goal] calorie goal	Move Ring Today [Time]

semantic differentials while also exploring the potential for new overall user experience factors could enhance the scale's utility and effectiveness.

6.5 Limitations

Our study has three main limitations. First, its online nature restricts the generalizability of findings to real-world VA interactions, which can be potentially influenced by contextual factors and day-to-day use. Such examples include asking a personal health data question during a walk or conversely while sitting at home relaxing. Thus, future research should extend our findings through in lab and real-world settings for enhanced validity and reliability. Second, our participant pool consisted of individuals familiar with VAs and personal health data (i.e., potential sampling bias) While we are confident that focusing initially on this demographic provides insights into general results, we acknowledge it may not fully represent the broader population for whom VAs could be used. To address this, future research could include participants with varying levels of VA familiarity and increasingly diverse demographic backgrounds (e.g., older adults). Third, while Study 2 focused on insight categories, only Value-based responses were utilized. While this aligned with the study's goal, it does leave other combinations of insight category and response type to still be evaluated which in turn could provide increasingly fine-tuned guidelines for VA answers. Furthermore, while our study offered heart rate and step count as data types of choice throughout each study, to engage participants within the study, we recognize that questions pertaining to specific data types could result in different desired answers structures. Future work could perform a comparative study across the many data types captured within one's personal health data. Our studies offer initial insights into answer structures for VA interactions involving personal health data questions. By exploring diverse response types and insight categories, notably applicable to any data type, we lay the groundwork for designing and developing VA interactions involving personal health data.

7 CONCLUSION

Through the use of a custom-built browser-based pseudo-voice assistant (VA), our work investigates differing answer structures in response to personal health data queries. Two user studies involving a total of 82 participants were conducted, during which participants interacted with our VA, posing questions and ranking their experiences and preferences of three distinct answer structures: minimal, keyword, and full sentence. We provide empirical findings that reveal a notable preference for full sentence answers, which consistently demonstrated higher quality, behavior, comprehensibility, and efficiency across various response types (Open, Range, Binary, and Value) and personal health insight categories (Contextual, Preemptive and Proactive, Goal and Performance, Combination and Comparison, Historical and Trend, and Current Status). These results come at a contrast to previous work which explore answer structures for general VA use. Our results suggest that full sentence answers offer less ambiguity, and despite their longer response time, full sentence answers were perceived as equally efficient. Along with other findings, such as a desire for VAs to be efficient and technical rather than social entities (e.g., as a fitness coach), we provide design implications in line with these results that offer insight into future VA systems handling personal health data queries.

ACKNOWLEDGMENTS

We would like to thank all of our study participants for their time and effort involved in participation. The work of Pourang Irani was supported in part by an NSERC Discovery Grant on In-Situ User Interfaces, from which Bradley Rey was partly funded.

REFERENCES

[1] Emily Couvillon Alagha and Rachel Renee Helbing. 2019. Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri. *BMJ health & care informatics* 26, 1 (2019), e100075.

[2] Fereshteh Amini, Khalad Hasan, Andrea Bunt, and Pourang Irani. 2017. Data representations for in-situ exploration of health and fitness data. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/3154862.3154879

[3] Daniel L Ashbrook. 2010. *Enabling mobile microinteractions.* Georgia Institute of Technology.

[4] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310432

[5] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[6] Robin Brewer, Casey Pierce, Pooja Upadhyay, and Leeseul Park. 2022. An Empirical Study of Older Adult's Voice Assistant Use for Health Information Seeking. *ACM Trans. Interact. Intell. Syst.* 12, 2, Article 13 (jul 2022), 32 pages. https://doi.org/10.1145/3484507

[7] Stephen Brewster, Joanna Lumsden, Marek Bell, Malcolm Hall, and Stuart Tasker. 2003. Multimodal 'eyes-free' interaction techniques for wearable devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 473–480. https://doi.org/10.1145/642611.642694

[8] Eugene Cho, S. Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. 2020. Will Deleting History Make Alexa More Trustworthy? Effects of Privacy and Content Customization on User Experience of Smart Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376551

[9] Eun Kyoung Choe, Bongshin Lee, and m.c. schraefel. 2015. Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Computer Graphics and Applications* 35, 4 (2015), 28–37. https://doi.org/10.1109/MCG.2015.51

[10] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[11] Arlene E Chung, Ashley C Griffin, Dasha Selezneva, and David Gotz. 2018. Health and fitness apps for hands-free voice-activated assistants: content analysis. *JMIR mHealth and uHealth* 6, 9 (2018), e9705.

[12] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.

[13] Geoff Cumming and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *American psychologist* 60, 2 (2005), 170.

[14] Caroline de Cock, Madison Milne-Ives, Michelle Helena van Velthoven, Abrar Alturkistani, Ching Lam, and Edward Meinert. 2020. Effectiveness of Conversational Agents (Virtual Assistants) in Health Care: Protocol for a Systematic Review. *JMIR Res Protoc* 9, 3 (9 Mar 2020), e16934. https://doi.org/10.2196/16934

[15] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[16] Pierre Dragicevic. 2016. *Fair Statistical Communication in HCI*. Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13

[17] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 23, 12 pages. https://doi.org/10.1145/3469595.3469618

[18] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467.

[19] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[20] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep it Short: A Comparison of Voice Assistants' Response Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 321, 12 pages. https://doi.org/10.1145/3491102.3517684

[21] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Construction of UEQ+ scales for voice quality: measuring user experience quality of voice interaction. In *Proceedings of Mensch Und Computer 2020* (Magdeburg, Germany) *(MuC '20)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3404983.3410003

[22] Rafal Kocielnik, Raina Langevin, James S. George, Shota Akenaga, Amelia Wang, Darwin P. Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T. Hsieh, Kabir Yadav, Herbert Duber, Gary Hsieh, and Andrea L. Hartzler. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3469595.3469599

[23] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (07 2018), 1248–1258. https://doi.org/10.1093/jamia/ocy072 arXiv:https://academic.oup.com/jamia/article-pdf/25/9/1248/34150600/ocy072.pdf

[24] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/2030112.2030166

[25] Gesa Alena Linnemann and Regina Jucks. 2018. 'can i trust the spoken dialogue system because it uses the same words as i do?'—influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers* 30, 3 (2018), 173–186.

[26] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[27] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[28] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. https://doi.org/10.1145/3469595.3469600

[29] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications* 129 (2019), 56–67. https://doi.org/10.1016/j.eswa.2019.03.054

[30] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[31] Jum Nunnally. 1994. Psychometric theory. *(No Title)* (1994).

[32] Kenneth Olmstead. 2017. Nearly half of Americans use digital voice assistants, mostly on their smartphones. *Pew Research Center* 12 (2017).

[33] Adam Palanica, Anirudh Thommandram, Andrew Lee, Michael Li, and Yan Fossat. 2019. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *NPJ digital medicine* 2, 1 (2019), 55.

[34] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[35] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (<conf-loc>, <city>Denver</city>, <state>Colorado</state>, <country>USA</country>, </conf-loc>) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[36] Pallavi Rao Gadahad and Anirudha Joshi. 2022. Wearable Activity Trackers in Managing Routine Health and Fitness of Indian Older Adults: Exploring Barriers to Usage. In *Nordic Human-Computer Interaction Conference* (Aarhus, Denmark) *(NordiCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 7, 11 pages. https://doi.org/10.1145/3546155.3546645

[37] Bradley Rey, Charles-Olivier Dufresne-Camaro, and Pourang Irani. 2023. Towards Efficient Interaction for Personal Health Data Queries on Smartwatches. In *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction* (Athens, Greece) *(MobileHCI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 18, 7 pages. https://doi.org/10.1145/3565066.3608700

[38] Bradley Rey, Bongshin Lee, Eun Kyoung Choe, and Pourang Irani. 2023. Investigating In-Situ Personal Health Data Queries on Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 179 (jan 2023), 19 pages.

https://doi.org/10.1145/3569481

[39] Bradley Rey, Bongshin Lee, Eun Kyoung Choe, and Pourang Irani. 2024. Databiting: Lightweight, Transient, and Insight Rich Exploration of Personal Data. *IEEE Computer Graphics and Applications* 44, 2 (2024), 65–72. https://doi.org/10.1109/MCG.2024.3353888

[40] Lionel Robert. 2017. The growing problem of humanizing robots. *Robert, LP (2017). The Growing Problem of Humanizing Robots, International Robotics & Automation Journal* 3, 1 (2017). https://doi.org/10.15406/iratj.2017.03.00043

[41] Martin Schrepp and Jörg Thomaschewski. 2019. Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence* 5, 7 (12/2019 2019), 88–95. https://doi.org/10.9781/ijimai.2019.06.006

[42] Jaisie Sin, Dongqing Chen, Jalena G Threatt, Anna Gorham, and Cosmin Munteanu. 2022. Does Alexa Live Up to the Hype? Contrasting Expectations from Mass Media Narratives and Older Adults' Hands-on Experiences of Voice Interfaces. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–9.

[43] 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*. 1–13.

[44] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Noé Zufferey, Mathias Humbert, Kévin Huguenin, and Mauro Cherubini. 2022. Are Those Steps Worth Your Privacy? Fitness-Tracker Users' Perceptions of Privacy and Utility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 181 (dec 2022), 41 pages. https://doi.org/10.1145/3494960

[45] Peter West, Max Van Kleek, Richard Giordano, Mark J. Weal, and Nigel Shadbolt. 2018. Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174058

[46] Jayme L Wilder, Devin Nadar, Nitin Gujral, Benjamin Ortiz, Robert Stevens, Faye Holder-Niles, John Lee, and Jonathan M Gaffin. 2019. Pediatrician attitudes toward digital voice assistant technology use in clinical practice. *Applied clinical informatics* 10, 02 (2019), 286–294.

[47] Michael Zimmer, Priya Kumar, Jessica Vitak, Yuting Liao, and Katie Chamberlain Kritikos. 2020. 'There's nothing really they can do with this information': unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society* 23, 7 (2020), 1020–1037. https://doi.org/10.1080/1369118X.2018.1543442 arXiv:https://doi.org/10.1080/1369118X.2018.1543442