



Can we talk?

Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent

Maharjan, Raju; Rohani, Darius Adam; Bækgaard, Per; Bardram, Jakob; Doherty, Kevin

Published in:

CUI '21: CUI 2021 - 3rd Conference on Conversational User Interfaces

Link to article, DOI:

[10.1145/3469595.3469600](https://doi.org/10.1145/3469595.3469600)

Publication date:

2021

Document Version

Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):

Maharjan, R., Rohani, D. A., Bækgaard, P., Bardram, J., & Doherty, K. (2021). Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *CUI '21: CUI 2021 - 3rd Conference on Conversational User Interfaces* [5] Association for Computing Machinery. ACM International Conference Proceeding Series <https://doi.org/10.1145/3469595.3469600>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent

Raju Maharjan
Technical University of Denmark
Kgs. Lyngby, Denmark
rajm@dtu.dk

Darius Adam Rohani
Technical University of Denmark
Kgs. Lyngby, Denmark
daroh@dtu.dk

Per Bækgaard
Technical University of Denmark
Kgs. Lyngby, Denmark
pgba@dtu.dk

Jakob E. Bardram
Technical University of Denmark
Kgs. Lyngby, Denmark
jakba@dtu.dk

Kevin Doherty
Technical University of Denmark
Kgs. Lyngby, Denmark
kevdoh@dtu.dk

ABSTRACT

The growing popularity of smart-speakers in recent years has led to increased interest in the capacity of Conversational Agents (CAs) to support health and wellbeing. This extends to their potential to engage users in human-like conversations as means of gathering self-reported health data. Prior research has focused on the optimization of CAs for the collection of discrete responses to standardized questionnaires. Less research however, has investigated how a more conversational modality shapes what people recount of their wellbeing nor what they make of the experience. This paper presents the findings of a lab-based random assignment study contrasting 59 participants' experiences of two distinct designs of a CA named Sofia — each separately enabling discrete or open-ended responses to the World Health Organization-Five Wellbeing Index (WHO-5) questionnaire. Analysis of task completion times, Speech-System Interface Usability (SASSI) scores, and coherence between verbal and paper-based responses suggests that CAs can serve as a feasible means of gathering self-reported health data, although users report finding discrete response options more habitable (i.e. easier to grasp) than an open-ended alternative. We discuss the implications of these findings for the design of CAs to support the self-report of health and wellbeing, and highlight future research directions.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; *Laboratory experiments*.

KEYWORDS

conversational agent; voice user interface; conversational user interface; mental health; wellbeing; self-report; WHO-5; SASSI

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CUI '21, July 27–29, 2021, Bilbao (online), Spain

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8998-3/21/07...\$15.00

<https://doi.org/10.1145/3469595.3469600>

ACM Reference Format:

Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob E. Bardram, and Kevin Doherty. 2021. Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3469595.3469600>

1 INTRODUCTION

Self-reports obtained via questionnaires such as the World Health Organization-Five Well-Being Index (WHO-5) and Patient Health Questionnaire (PHQ-9) are regarded as a valuable source of insight into peoples' subjective experiences, and an effective means of monitoring and assessing mental health and wellbeing [51]. Traditionally carried out using pen and paper, these questionnaires are increasingly administered via smartphones and myriad other graphical user interface (GUI) devices [17, 53]. Although these systems offer advantages in terms of ease of use and access, research points to the user burden associated with their use as a primary limitation of their user experience [47], and particularly so for open-ended questionnaires given the additional time and effort required to formulate a response [8]. Such findings surface the need to attend to the design, and limitations, of a variety of self-report technologies in order to realize their potential to support understanding and care.

Indeed, self-report plays an essential role in our understanding of mental health and illness, yet itself comprises myriad forms of self-expression and communication [17, 47]. Recognizing the close-knit relationship between mental health knowledge and self-expression therefore compels us to consider the value of alternative interaction media. Recent advancements in speech-enabled interfaces (e.g., Amazon Alexa and Google Assistant) suggest the potential of voice enabled Conversational Agent (CA) technologies, to enable more 'natural', 'human' and engaging forms of self-expression, and perhaps in turn more accurate, honest and insightful disclosures of emotional experience [16, 37]. The question arises, in addition to yielding advantages for users with visual and motor impairments [15] and despite outstanding technical limitations including an 8 to 12 second limit on user responses [29], might CAs serve as alternative, accessible and engaging avenues for self-expression?

The growing trend in ownership of smart speaker devices including Amazon's Echo and Google's Home [46] has led to increased interest in these systems' potential to support healthcare [25, 27, 28, 31, 44]. Fifty-two percent of respondents to a recent survey of U.S. adults ($n = 1,004$) reported an interest in the use of Virtual Assistants (VAs) for healthcare, while 7.5% reported having made use of such systems for a healthcare-related task such as inquiring about symptoms of illness, searching for information concerning medication use, and seeking care and treatment options [5].

Research within Human-Computer Interaction (HCI) to date has primarily focused on issues of privacy, usability, user satisfaction and attitudes towards VAs [9, 13, 15, 32]. Recent work has, for example, focused on the impact of CAs' voice characteristics on users' experiences [6, 10, 18], and the effectiveness of chat-based surveys compared to traditional web-based methods [7, 55]. Others have proposed initial guidelines for the design [41, 50, 54] and evaluation of CAs [26]. Studies in the healthcare domain have likewise examined technical performance, user experience [4, 16], and health-related outcomes including perceptions of therapeutic alliance, trust, and human intervention [25] [28, Table 3].

Despite these initial research efforts however, and in light of the growing potential of these systems, there remains much we do not know about the feasibility and design of CAs for the self-report of mental health and wellbeing. While initial research efforts suggest that voice-based self-report technologies, including Interactive Voice Response (IVR) [2] and Automatic Speech Recognition (ASR) [33], can serve as efficient tools for home-based health monitoring, most such systems have been designed to facilitate discrete responses to closed-ended questions (e.g., [23, 24, 34]).

In the same way that many pen-and-paper-based questionnaires have been converted to mobile and web platforms for improved efficiency and accessibility in recent years, we are likely to soon see increased conversion of these now GUI-based questionnaires for delivery via CA. The direct adoption of GUI-centric models of interaction would however adversely impact the CA medium's potential to support more 'natural', human-human-like conversational interactions. Although significant future validation studies would be essential to the ethical adoption of these systems in clinical practice, it is therefore also important to foster an understanding of the relationship between different conversational design strategies (discrete and open-ended) and users' reporting behaviors and experiences to inform future CA designs.

This paper aims to bridge this gap by exploring the relationship between conversational design and users' self-reporting practices in the case of a CA for the self-report of mental health and wellbeing. Focusing on questions of both technological feasibility and user experience, we present a comparative analysis of 59 participants' responses to the WHO-5 on paper, and via two distinct designs of a CA named *Sofia*; the first requiring discrete responses and the second allowing for open-ended answers to the same series of questions. We furthermore compare participants' CA self-reporting experiences as measured by task completion times, Subjective Assessment of Speech-system Interface Usability (SASSI) questionnaire and participants' comments in-study. Adopting a mixed-methods approach, we therefore seek to inform the future design of CAs for health and wellbeing by addressing three research questions, exploring;

RQ1: Feasibility & Coherence. To what extent do users' responses to a paper-based wellbeing scale align with those provided to a CA?

RQ2: Behavior & Experience. How are users' self-report behaviors and experiences shaped by distinct (discrete vs. open-ended) conversational designs?

RQ3: Implications for Design. What in turn are the implications of these conversational dialog choices for the design of CAs to support the self-report of health and wellbeing?

The results of this study suggest that CAs can serve as a viable medium for the self-report of health and wellbeing. Participants provided faster responses to a wellbeing scale delivered via CA and eliciting open-ended responses, yet deemed a design of the same CA seeking discrete responses as more habitable — defined as the extent to which the user is aware of what to say and what the system is doing [21] — in nature. Based on interpretation of these findings, this paper concludes with a discussion of the feasibility of CAs for the self-report of health and wellbeing as well as reflection on the value of a voice-based approach to self-report, and an increasingly conversational approach to CA design.

By examining the impact of conversational design choices on self-reporting practice and experience, this study (i) demonstrates the feasibility of a CA-based approach to the self-report of wellbeing in terms of coherence with a standardized paper-based scale, (ii) elucidates meaningful differences between discrete and open-ended CA-based self-report experiences in terms of user response times, help invocations, fall-back intents, and responses to the SASSI user experience scale, and finally (iii) offers implications for the design of conversational user interfaces to support the self-report of health and wellbeing.

2 RELATED WORK

This study builds upon and contributes to two increasingly overlapping bodies of literature; those concerning self-report technologies and conversational interaction design.

2.1 Conversational Agents for Self-Report

Prior research has examined the design and use of voice-based self-report technologies in a variety of healthcare contexts [19, 31]. Both IVR and ASR technologies have been successfully deployed for the in-home monitoring of patients with chronic diseases for example. Azzini et al. developed a prototype telephone-based dialog system enabling hypertensive patients to record pertinent health data by calling a toll-free number, and therefore mitigating the need for a clinical visit [2]. Levin et al. conducted a usability evaluation of the ASR-based 'Pain Monitoring Voice Diary' system, finding users able to navigate the flexible interface, and that self-reporting efficiency increased with users' experience, both in terms of session duration and avoidance of troublesome dialog scenarios [33].

While such studies have served to provide valuable insight into the design of dialog systems, recent advancements in Natural Language Processing (NLP) and VA interfaces have made more conversational forms of interaction possible. One such CA, implemented within Alexa and titled 'Symptom Checker' [23], aims to support users in discovering possible causes of their reported symptoms

by answering a series of questions. ‘Wellness Guru’ [24] and ‘Depression Test’ [34] strive to produce assessments of users’ mental health by posing questions from mental health questionnaires such as the PHQ-9.

While these particular CA applications are designed to elicit discrete responses to questions, voice technology is not limited to such forms of input. DeVault et al. for example, designed a virtual human interviewer to assess mental health conditions including depression and anxiety via automated analysis of verbal and non-verbal behaviors [16]. Results from this user study suggested that the system was able to engage users in an open-ended conversation as long as 15 to 25 minutes in duration, and that participants were comfortable sharing intimate information. Kocielnik et al. designed a CA named ‘Robota’ that asked users to provide ten open-ended daily reflections. Comparing speech- and text-based interactions, the authors found that voice interaction enabled users to step back and reflect on their work as well as opportunities for workplace-related behavior-change [29].

These studies suggest that speech-based CAs represent a unique opportunity to serve as alternative media for the self-report of health and wellbeing, in turn fostering more honest, accurate and insightful forms of self-expression. However, while Voice User Interfaces (VUIs) are seeing increasingly widespread adoption among consumers, the research literature remains replete with findings and commentary reflecting the limitations of these technologies; from speech recognition errors to unintelligible responses and difficult to navigate dialog flows [36, 48, 50]. This brings us to the question of design.

2.2 Conversational Interaction Design

The potential of CA technologies hinges upon the design of efficient, engaging and effective conversations [38]. Research furthermore points to appropriate interaction design as essential to complementing speech processing in ways that compensate for the less-than-perfect accuracy of these systems [1, 39, 40]. And yet, there exists a general consensus among both HCI and healthcare researchers that limited design-oriented work has been conducted concerning CAs to date [12, 25, 28].

In 2003, Bernhard Suhm proposed a set of ten relevant VUI design guidelines for a telephone-based dialog system [50, Table 3]; including to (i) keep it simple, (ii) carefully control the amount of spoken output, (iii) provide options which match the ways users think, (iv) minimize the acoustic confusability of vocabulary, (v) provide carefully designed feedback, (vi) abide by natural turn-taking protocols, (vii) coach a little at a time, (viii) offer alternative input modalities, (ix) that yes/no queries can be very robust, and to (x) carefully select the appropriate persona. Wei and Landay recently extended these guidelines to include those provided by Google and Amazon, resulting in 5 categories of 17 Speech User Interface (SUI) heuristics [54]. Murad et al., in summarizing much of this work, suggest that the path towards more conversational voice-based interactions must in turn be based on VUI heuristics which build upon existing GUI design principles [41].

Other researchers have conducted evaluations of more specific conversational design features. Motalebi et al., for example, conducted a pilot study to explore the impact of random back-channeling

(i.e., “hm”, “uhum”, “aha”, and “yeah”) on user engagement [37]. While most participants in this study were open to receiving such verbal cues, many suggested providing more appropriately tailored responses in addition to back-channeling to improve active listening performance. Myers et al. identified unfamiliar intents, NLP errors, failed feedback and system errors as the primary obstacles encountered by users while creating, modifying, deleting, and inviting attendees to a voice-based calendar titled ‘DiscoverCal’ [42]. Guessing was furthermore identified as the tactic most commonly employed by users to overcome such obstacles. Cambre et al. conducted an online survey to evaluate a diverse selection of synthesized and human voices according to the user’s listening experience, as well as perceptions of clarity, quality and comprehension. The authors concluded that while synthesized voices can come close to human voices, no single voice outperformed all others across all evaluation dimensions [6]. Similarly, Dubiel et al. evaluated the impact of different synthetic voices on participants’ perceptions and behavior during a flight-booking task conducted via CA, concluding that although users perceived significant differences between synthetic voices in terms of truthfulness and engagement, this did not translate into significant differences in behavior [18]. Choi et al. conducted a 20-day in-home study to investigate visually impaired people’s CA use and perceptions of different speech rates [10]. They found that visually impaired users were generally more satisfied with a default human rate of speech, and emphasized speech rate control as an important feature of CA design. Chu et al. in 2005 proposed a simple taxonomy positing three primary variants of interaction; (i) Finite-state, in which case the CA asks the user specific questions and assesses the user’s input as valid or otherwise; (ii) Frame-based, in which a series of slots are filled with data captured from the user; and (iii) Free-form, in which case users’ input is accepted without restriction [11]. This final dialog management strategy has more recently been additionally described as ‘Agent-based’ [35].

Whilst these research efforts bring us closer to an understanding of the underlying features of effective CA design, their nascent form also highlights the need for a richer understanding of users’ experiences of interaction with CAs as shaped by diverse dialog management strategies [36]. We therefore undertake, in this paper, an attempt to generate a richer understanding of the relationship between distinct approaches to conversational design and users’ reporting practices and experiences.

3 METHOD

A mixed-methods study was designed to enable analysis of (i) the feasibility of CAs for the self-report of health and wellbeing according to the WHO-5 questionnaire in comparison to paper-based responses, and (ii) to generate an understanding of the impact of distinct conversational designs on users’ reporting behaviors and experiences.

3.1 Participants

Considering the emergent nature of CAs and young people’s willingness to engage with new technologies, sixty participants (F=15, M=45) were recruited from a local university via internal email, posters, and word of mouth. Participants had a mean age of 26.33

Table 1: Sample conversations between a participant and both designs of Sofia

Design A: Discrete	Design B: Open-ended
User: OK Google, talk to <i>Sofia</i> .	User: OK Google, talk to <i>Sofia</i> .
<i>Sofia</i> : Hi, I'm <i>Sofia</i> . Please answer each of the five questions on the scale of 0 to 5. 0 being at no time, and 5 being all of the time. Over the last 2 weeks, how often have you felt cheerful and in good spirits?	<i>Sofia</i> : Hi, I'm <i>Sofia</i> . Please answer each of the five questions using a word or a phrase. Over the last 2 weeks, how often have you felt cheerful and in good spirits?
User: One	User: Often...definitely.
<i>Sofia</i> : How often have you felt calm and relaxed?	<i>Sofia</i> : How often have you felt calm and relaxed?
User: Two	User: I have felt surprisingly calm and relaxed recently.
<i>Sofia</i> : How often have you felt active and vigorous?	<i>Sofia</i> : How often have you felt active and vigorous?
User: Four	User: I have not felt vigorous very often.
<i>Sofia</i> : How often have you woken up feeling fresh and rested?	<i>Sofia</i> : How often have you woken up feeling fresh and rested?
User: Three	User: Everyday.
<i>Sofia</i> : How often have you felt that your daily life has been filled with things that interest you?	<i>Sofia</i> : How often have you felt that your daily life has been filled with things that interest you?
User: Five	User: Also everyday...I'd say.
<i>Sofia</i> : Thank you for answering the questions. Bye!	<i>Sofia</i> : Thank you for answering the questions. Bye!

(SD = 4.32) years and consisted of university staff and students across multiple disciplines. All participants but one reported that they were proficient English speakers. Fifty-eight percent of participants reported that they had no prior experience using CAs. Participants were offered refreshments during their participation, and the opportunity to win a Google Home Mini device by entering their names into a lottery drawn at the end of the study.

3.2 Apparatus

We developed *Sofia* using Dialogflow¹ and deployed this CA via Google Home. The *Sofia* dialog flow integrates the five questions of the WHO-5 wellbeing scale [3], chosen as it asks simple questions amenable to a variety of forms of self-report, is less invasive than many clinical scales, and is appropriate for a general population group [51]. Slight variations were incorporated into the wording of the preamble and the questions in line with prior digital mental health research practices [52], and in order to render the questionnaire more conversational in nature, as shown in Table 1. *Sofia* provides fallback re-prompts following two types of error: (i) 'no response' when the respondent takes too long to respond; and (ii) 'no match' when *Sofia* fails to understand the response. Respondents have three attempts to respond to each question. After three re-prompts, *Sofia* ends the conversation. Respondents can also end the conversation at any time by voicing the phrase "quit" or "stop". If needed, users can ask *Sofia* to repeat the question by stating "repeat" or "what was the question?", and can also ask for help by stating "help" or "what are my options," in which case *Sofia* repeats the preamble ("You can answer the question on the scale of 0 to 5; 0 being at no time, and 5 being all of the time"). On completion of the questionnaire, *Sofia* thanks the user for answering the questions and ends with "goodbye".

Two versions of *Sofia* were designed to employ the WHO-5 questionnaire;

- *Design A: Discrete* — This finite-state-based design [11] of *Sofia* instructs respondents to respond to each question of the WHO-5 questionnaire on a scale ranging from 0 to 5, and accepts only numerical responses, as illustrated by the example shown in the left-hand side of Table 1.
- *Design B: Open-Ended* — This free-form design [11] of *Sofia* instructs respondents to respond to the questionnaire using words or phrases of their own choosing and without constraint. Participants were requested to keep their open-ended responses brief in order to minimize interruptions imposed by the technological limitation of an 8 to 12 second response time. As shown in the right-hand side of Table 1, users' responses can take any form, and 'no match' errors therefore do not apply.

3.3 Experimental Design & Procedure

We conducted a lab study consisting of a within-group experimental setup counterbalanced by a 2x2 Latin square design. To mitigate the possibility of participants' familiarity with the questionnaire in one form influencing their responses to the second, we randomized the order of the WHO-5 questions posed via CA and adopted a counterbalancing approach enabling us to assess the existence of such biases. Figure 1 provides an overview of the experimental design which consisted of the following phases;

3.3.1 Pre-study. Prior to the study, participants were asked to sign a consent form and instructed how to interact with *Sofia*. Instructions included how to invoke *Sofia*, ask for help or repeat the question, and end the conversation. They were also made aware of the visual cue provided by the Google Home device (a rotating light) to infer when the CA is listening to and processing the participant's response. Participants were then asked to complete a demographic questionnaire containing questions concerning age, sex, occupation, proficiency in English, and prior experience using CAs.

¹<https://dialogflow.com>

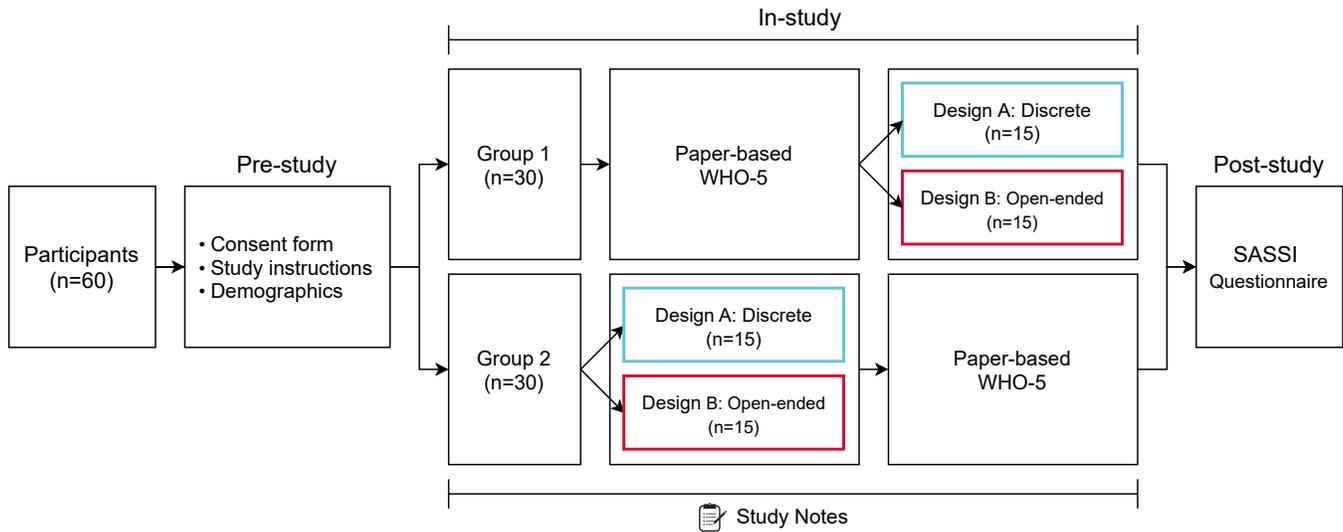


Figure 1: Experimental design and procedure. Participants were balanced in a 2x2 Latin square, such that half of the participants in each group were randomly assigned to the paper version either before (Group 1) or after (Group 2) *Sofia* designs.

3.3.2 In-study. During the study, half of the participants (Group 1) responded to the paper-based WHO-5 questionnaire followed by *Sofia*, and the other half (Group 2) responded to the questionnaire using *Sofia* first, followed by the paper scale. Participants were randomly assigned to one of the *Sofia* designs such that half of the participants in each group interacted with design A and the other half with design B. We asked participants to complete the WHO-5 questionnaire on paper in order to enable us to compare their responses to those provided through *Sofia*. This would allow us to assess the feasibility of each CA design as means for the self-report of wellbeing according to the WHO-5 scale.

Data collected through *Sofia* included participants' responses to the WHO-5 questionnaire (automatically transcribed to a string), timestamps for each question and response, as well as the number of times 'help', 'repeat' and 'fallback' intents were invoked. Participants' informal comments, suggestions, and other feedback during their interactions with *Sofia* were also noted.

3.3.3 Post-study. Upon completion of the study, participants filled out the SASSI questionnaire, providing their own subjective assessment of each *Sofia* design. This 34-item scale covers six user experience dimensions; system response accuracy, likability, cognitive demand, annoyance, habitability, and speed [22].

3.4 Data Processing

Both the SASSI questionnaire responses and data collected from *Sofia* were analyzed in R (v. 3.5.2). Three separate analyses were conducted for each dependent variable, namely, task completion time, SASSI scores, and the correlation between both *Sofia* designs and the paper version of the WHO-5 questionnaire.

Total task completion time was calculated as the difference between response initiation and registration times (from the point

at which the agent finished asking the question to that at which the participant completed their response) in *Sofia*. Participants' responses to the SASSI questionnaire according to each experience dimension were summed, taking each question's positive or negative directionality into account [21, Table 4].

To enable analysis of the correlation between participants' responses to both *Sofia* designs and on paper, three authors independently classified users' open-ended responses according to the response categories of the WHO-5 questionnaire. For example, participant P8's open-ended response "every day" was mapped to the number "5" and "in the weekends" to the number "2". There was substantial inter-rater agreement among classifiers, as assessed by Fleiss' kappa ($\kappa = 0.68$) [30]. Disagreements (43/134, 27%) were discussed and resolved. A small number of responses (2/145, 1.4%) were not captured and stored on the server, while several others were transcribed too indistinctly to classify (9/145, 6.2%) (e.g., "Bob's 70s", "I would say") according to the standard response options. Participants with missing or indistinct responses were removed from the correlation analysis.

Two authors discussed the notes taken by the observer during the experimental procedure, including all comments made by participants, and employed an affinity-diagramming approach to provide qualitative context for our key quantitative findings.

3.5 Statistical Analysis

In the case of all statistical analyses conducted, normality was assessed for each dependent variable by inspecting the distribution of the residuals and visually comparing the outer quantiles with a standard normal distribution. In all cases, the residuals did not conform to a normal distribution. Consequently, we employed non-parametric Mann-Whitney U test [43] statistics to compare both designs. The assumptions for a Mann-Whitney U test were carefully

Table 2: Summary statistics

	Discrete	Open-ended
# Participants (% female)	30 (23%)	29 (27%)
Age (M, SD)	26.63, 3.79	26.07, 4.93
Prior CA experience (%)	12 (40%)	12 (41%)
Total duration in seconds (M, SD)	68.31, 19.28	57.62, 10.99
CA response correlation with WHO-5 paper	.91	.77
# Help	6	1
# Repeat	8	2
# Fallback	19	0

assessed and met; including ensuring a sufficient degree of similarity between dependent variable distributions.

We chose to remove outliers; defined as any value more than 2.5 SD from the global average. This resulted in the removal of participant P50 from the task completion time metric (199 sec, global average = 62.77 sec), and P31 from SASSI questionnaire scores (M = 3.26, global average = 4.57).

These analyses enabled investigation of both the feasibility and experience of two distinct CA designs for self-reporting practice; assessed in terms of task completion times, SASSI scores, and the correlation between participants' responses to the WHO-5 questionnaire via both *Sofia* designs with the same participants' responses on paper.

4 RESULTS

We report results from a lab study conducted with 59 participants². Table 2 presents a summary of the data collected and subsequent statistical analyses for each design.

4.1 Counterbalancing

To identify any possible bias in participants' responses between Groups 1 and 2, we calculated the carry-over group effect for participants' overall WHO-5 scores and task completion times. We calculated the response difference as the Root-mean-square error (RMSE) between the WHO-5 scores on paper and as recorded through *Sofia* and did not find any significant group effect ($U = 303.0, p = .67$) between Group 1 (M = 4.44, SD = 4.48) and Group 2 (M = 5.33, SD = 5.84). Similarly, the task completion time between participants in Group 1 (M = 60.76, SD = 14.97 sec) was comparable with that of participants in Group 2 (M = 64.86, SD = 18.07 sec) in both *Sofia* designs, indicating that the time to complete the WHO-5 questionnaire in either group did not differ to a statistically significant extent ($U = 416.5, p = .30$). We were therefore able to conclude that the counterbalancing was effective in preventing bias in the ordering of WHO-5 responses via *Sofia* and on paper.

²Of the 60 initial participants, a single participant failed to complete the study, due to constraints on their own time.

4.2 Response Correlation Between Paper and Conversational Agent

Assessing the feasibility of each CA design as a means of self-report required comparison of participants' responses to the WHO-5 questionnaire as completed both through *Sofia* and on paper. Comparison of discrete CA responses with the paper form yielded a Pearson's correlation coefficient of $\rho = .91$ (95% CI 0.81 – 0.96), and open-ended CA responses with the paper form, a Pearson's correlation coefficient of $\rho = .77$ (95% CI 0.51 – 0.90). Discrete CA responses were, therefore, more strongly correlated with the paper scale. This finding is further illustrated in Figure 2, which plots participants' responses to the WHO-5 questionnaire on paper along the x-axis, and to the *Sofia* designs (discrete design in blue, open-ended design in red) along the y-axis. As such, each point represents a response to one of the five questions of the WHO-5 questionnaire.

4.3 Task Completion Time

We observed a significant difference ($U = 572.5, p < .01$) when comparing task completion times for each *Sofia* design. Participants providing open-ended responses took less time to complete the questionnaire (M = 57.62, SD = 10.99 sec) than those limited to discrete response options (M = 68.31, SD = 19.28 sec).

4.4 SASSI Scores

Figure 3 shows participants' average response score to each factor of the SASSI questionnaire in both *Sofia* designs. Participants reported significantly higher habitability in the case of the discrete response design ($U = 569.5, p = .02$) compared to the open-ended response design. However, there were no statistically significant differences between the two CA designs in terms of system response accuracy ($U = 371, p = .29$), likability ($U = 349.5, p = .19$), cognitive demand ($U = 396, p = .88$), annoyance ($U = 353, p = .40$), or speed ($U = 366, p = .52$).

4.5 Study Notes & Observations

The first author observed all participants' interactions with *Sofia*, and took notes reflecting participants' suggestions concerning possible improvements to the conversation design as well as feedback related to their personal experience of the CA. Suggestions included providing additional conversational cues to enhance engagement and avoid errors. P58, for example, suggested that “words and phrases such as ‘glad to hear that’, ‘sorry to hear that’ etc. would help me understand that it (*Sofia*) has actually understood and recorded the response...” and P34 stated that the “scale of the WHO-5 questionnaire is off (0-5)...” elaborating that they had trouble matching their response to a scale starting at zero rather than one. In terms of their experience of the practice of self-report via a CA, participants noted that they felt they had to be quite explicit in their responses, as P41 “when I talk to Google, I try to be specific” and P34 “I wanted to count the days and then answer the question” described. Finally, P39 commented that their initial expectations regarding the CA's capacity to answer questions were low; “It's not natural to ask for ‘Help’ to the device”. Participants' comments provide context for further reflection on the findings of this study, as we discuss next.

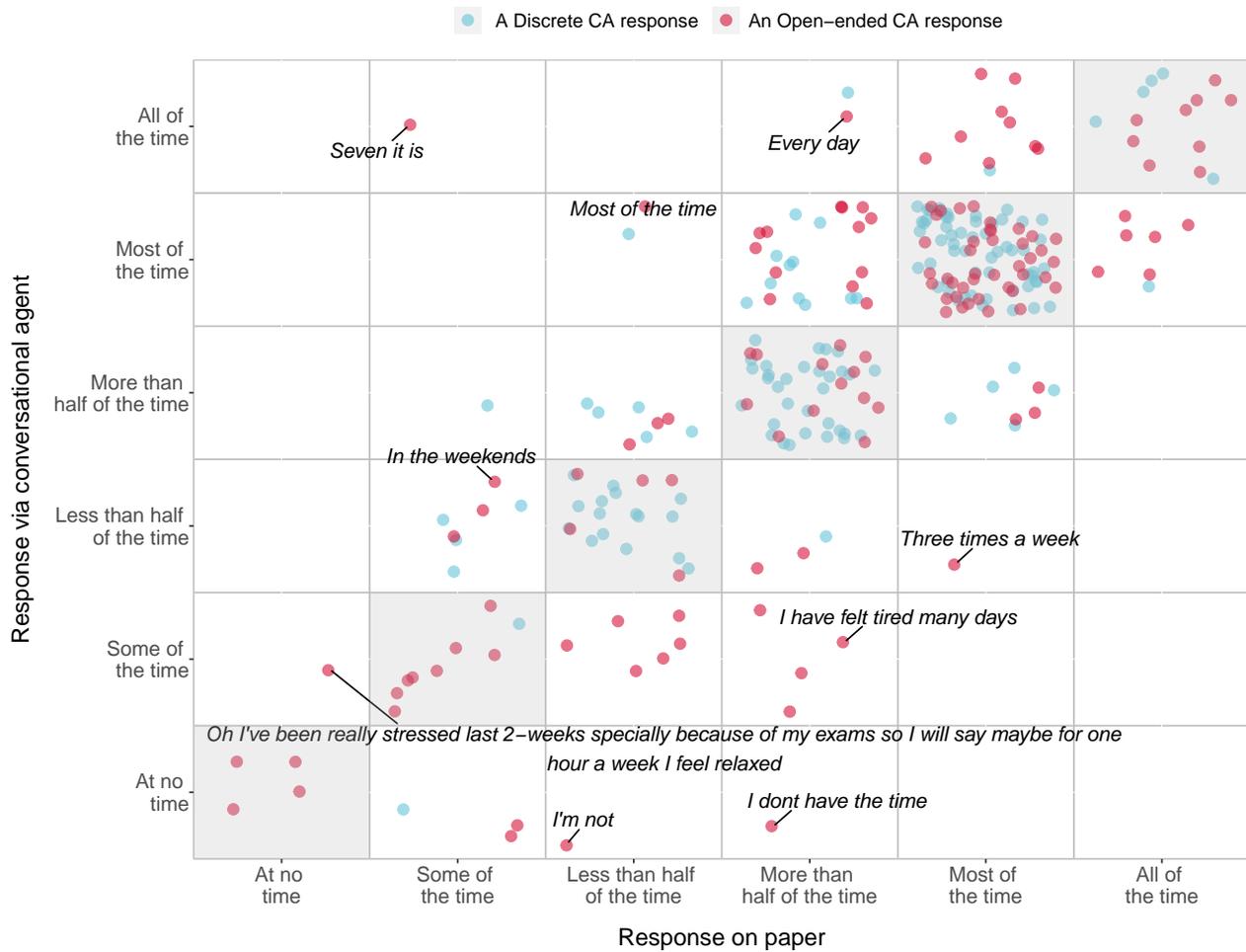


Figure 2: Coherence between responses to the WHO-5 questionnaire as provided through *Sofia* and on paper. Each dot represents a user’s response to a question. The x-coordinate represents a unique user’s response to each WHO-5 question on paper, and the y-coordinate that same user’s response to the same WHO-5 question as provided through *Sofia*. The 5 dots located in the far left column, for example, represent the 5 instances of “at no time” responses to the paper-based scale. The closer the responses are to the diagonal running from bottom left to top right, the more coherent (strongly correlated) they are. Of the 5 “at no time” responses for example, 4 gathered via CA matched exactly those provided on paper, as represented by the 4 dots in the bottom left square. Details of the raw transcription provide examples of the more incoherent open-ended responses. The exact coordinate positions of the scatter points are randomized within each square to facilitate clearer visual representation.

5 DISCUSSION

This work explored the potential of voice-based interactions for the realization of accurate, engaging and expressive self-report experiences. Our results suggest that CAs can serve as a feasible means of administering a standardized wellbeing scale, and that while permitting open-ended responses to a questionnaire in CA form has the potential to produce a more complete understanding of users’ health and wellbeing, discrete response options are deemed more habitable by users, meaning that users’ experiences of the system better aligned with their prior expectations of the technology’s capabilities. We reflect on these findings in light of our original research questions.

5.1 Feasibility & Coherence. To what extent do users’ responses to a paper-based wellbeing scale align with those provided to a CA?

This study reveals a strong correlation between participants’ responses to the WHO-5 scale on paper and through *Sofia*, suggesting that CAs can serve as a viable option for the administration of self-report questionnaires.

We did find a higher correlation between responses provided on paper and the discrete CA design as compared to the open-ended CA design. This may be explained in part by the fact that the discrete CA response options were identical to those of the paper version, whereas users of the open-ended CA design were

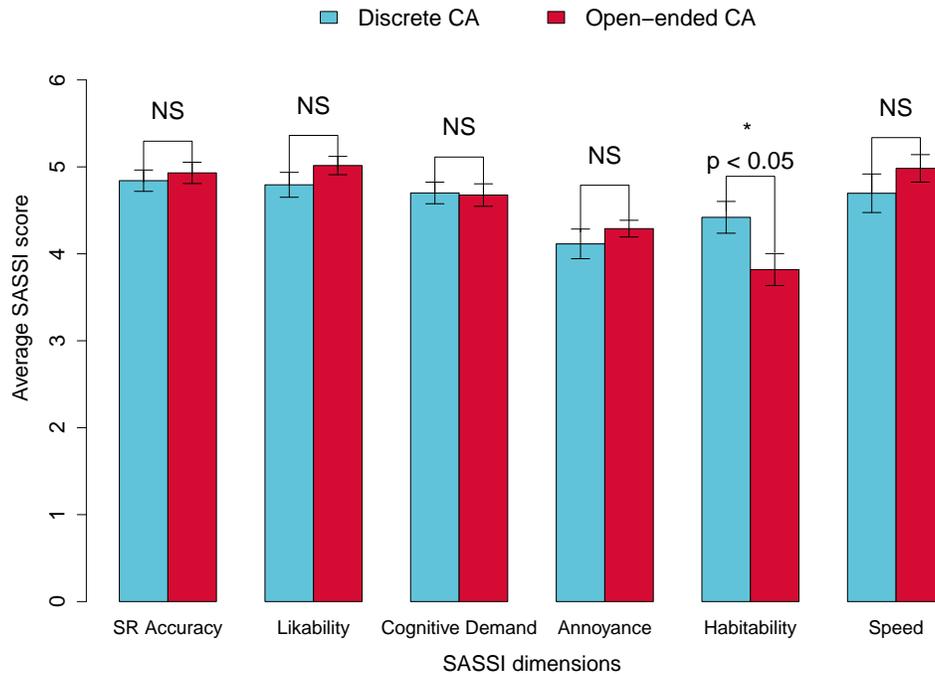


Figure 3: Participants’ responses to the SASSI questionnaire. Error bars reflect standard error and mean scores across all six dimensions. Note that the directionality of these constructs is based on the instructions provided alongside the scale i.e. the higher the score, the better the experience. For example, a higher score on the dimension of annoyance reflects a less annoying experience (SR Accuracy = System Response Accuracy, NS = Non-significant).

able to provide responses beyond the standard scale. For example, participants P15, P17, P18 and P58 responded by commenting “*half of the time*” – a response that does not map directly to any single option of the standard WHO-5 scale.

This finding of lower coherence between a paper-based scale and the open-ended CA condition may also be related to differences in users’ self-expression on paper and as enabled via CA. As shown in Figure 2, P8, for example, responded to the question “How often have you felt calm and relaxed?” by noting “*Every day*”, which was interpreted by human raters as representing “All of the time”, and yet the participant’s response to the same question on paper was “More than half of the time”. Additionally, we observed that although participants were asked to respond using a single word or phrase in the open-ended CA design, they often employed complete sentences. P27 and P23, for example, responded to one question by commenting “*I think most of the time because my course is not super stressful*”, and “*I have felt surprisingly calm ... relaxed recently*” respectively. This introduced additional challenges in terms of mapping users’ responses to the standardized WHO-5 scale, although it allowed users to think out loud, reflect on their experiences, and provide additional context for their comments.

Despite the complexities that open-ended reporting introduces when mapping users’ responses to a standardized scale, this approach did prove feasible in the majority of cases, suggesting that self-reports collected through an open-ended CA design may enable a richer understanding of respondents’ health and wellbeing. In

this study, users’ open-ended responses were manually mapped to a standardized scale, a practice which future studies could consider implementing via machine learning techniques and sentiment analysis (e.g., [14, 20, 45]). This may not only make for a more efficient process but enable the real time assessment of users’ open-ended responses, potentially allowing for and supporting additional in-situ reflection on their wellbeing. While such methods have been broadly employed in other scientific fields to support information-extraction from text-based corpora, extensive validation studies would be required prior to their adoption for clinical purposes.

While this comparison of two distinct CA designs suggests the feasibility of both approaches, it also communicates the value of questioning motivations for translating standardized scales to CA form, and reveals an opportunity for future research efforts to explore even more creative and novel forms of conversational design as a means of realizing the potential of these systems.

5.2 Behavior & Experience. How are users’ self-report behaviors and experiences shaped by distinct (discrete vs. open-ended) conversational designs?

Analysis of task completion times and participants’ responses to the SASSI questionnaire with respect to each *Sofia* design reveals that different CA response options (discrete vs. open-ended) significantly impact users’ self-reporting behaviors and experiences.

5.2.1 Users' Self-Report Behaviors. Results reveal that participants took longer to complete the WHO-5 questionnaire in the discrete CA rather than open-ended CA condition. This unexpected, and somewhat counter-intuitive, finding may be partially explained by the fact that participants invoked a higher number of 'help', 'repeat' and 'fallback' intents when providing discrete responses.

We observed that participants at times struggled to remember the response scale, which might have affected their cognitive capacity to pay attention to the question, resulting in more frequent invocation of the 'help' intent. Similarly, participants in the discrete CA design often took a long time to decide between the pre-defined response options, resulting in 'no-response' fallback re-prompts. It is also possible that participants' desired response was not present among the options provided by the system, necessitating increased reflection prior to responding. P34, for example, began to count days passed in order to enable them to accurately respond to each question, prolonging their response time and leading to the invocation of 'no-response' intents.

Hence, it could conceivably be hypothesized that providing pre-defined response options may not necessarily lead to a more efficient self-report experience for CA users. If not well-designed, this feature may slow down users' decision-making, introduce higher error rates, and require longer task-completion times.

5.2.2 Users' Self-Report Experiences. Our analysis of participants' responses to the SASSI questionnaire shows that the discrete CA design was perceived as significantly more habitable than that of the open-ended CA. Hone et al. define a habitable system as "one in which there is a good match between the users' conceptual model of the system and the actual system" [22]. Unlike in the case of the discrete CA, the open-ended CA did not provide precise response options to the questionnaire. This may have proved unexpected for some users and could also have generated uncertainty regarding what they should say or whether they were employing the correct terms.

The fact that habitability was the only dimension of the SASSI questionnaire on which the open-ended CA was viewed less favorably by users suggests that participants might not have expected the CA to be able to receive unconstrained responses. In their perception, a CA understanding of user input might have been limited to a pre-determined set of response options as in the case of the discrete CA. Clark et al. report a similar mismatch in their prior work, suggesting that open-ended and conversational forms of interaction may run counter to users' current perceptions of CAs' capabilities [13].

5.3 Implications for Design. What in turn are the implications of conversational dialog choices for the design of CAs to support the self-report of health and wellbeing?

As Stone and Shiffman have noted, "care in the design of questions and response formats can greatly reduce the burden by enabling minimal response time and by making the task more pleasant" [49]. In this instance, however, providing discrete response options did not lead to a more efficient reporting process for users; a finding which highlights the importance of attaining a rich understanding

of the practice of self-report in conversational interaction. Drawing on the results of this study and our observations of users' behaviors and experiences, we are able to suggest several implications for CA design to support the self-report of health and wellbeing.

5.3.1 Employ conversational cues to support engagement. While most participants made use of the visual cue (a bright light) provided by the Google Home device to infer when the CA was listening to and processing their conversation, a small number noted that they had not paid attention to this feature; commenting that phrases such as "glad to hear that" when recounting positive experiences, or "sorry to hear that" when responding negatively could make the conversation more engaging, and at the same time reflect that the CA has understood and recorded their response. We therefore, suggest integrating conversational cues into the design of CAs to administer self-report questionnaires — although designers must also then carefully navigate the increasingly thin distinction between automated assessment and intervention.

Sentiment analysis might furthermore be employed to support the provision of conversational cues by tailoring the CA's response in realtime to the emotional valence of users' comments; striking an empathetic tone following a negative user response, for example. Doing so may both enhance the conversational experience and reduce the conceptual gap between users' mental models of CA technologies and their actual interaction potentials. It is, however, also important to note that such design choices could entail the risk that a CA unintentionally provides incorrect conversational cues at the wrong times, leading to a negative user experience, and even potentially resulting in harm to users [4].

5.3.2 Avoid requiring users to guess the meaning of terms and response options. Guessing has been reported as the tactic most commonly employed by users to overcome obstacles in interacting with CAs [42]. In the context of health and wellbeing, such practices have the potential to yield significant adverse consequences, including the skewed interpretation of patients' wellbeing by health professionals. In this study, P46 and P48 did not understand the meaning of the word 'vigorous'. They asked the CA for its meaning, which it could not provide, and subsequently invoked a 'no-match' re-prompt. After twice attempting to obtain the meaning of the word from the CA, the participants responded to the question by guessing. When it comes to the use of such questionnaires within CAs, we therefore recommend keeping questions simple and providing the meaning of complex terms in order to avoid requiring users to guess the meaning of terms and response options.

5.3.3 'Help' proactively. Prior research has suggested that adopting GUI design guidelines, such as providing help and documentation, can be applied to certain aspects of VUI design, and help in identifying usability concerns [41]. In this study, relatively few (six) participants benefited from the 'help' option provided through *Sofia* although all were briefed about it. P33 forgot the scale according to which they were required to respond while using the discrete CA. Instead of asking for 'help', they tried to remember the relevant structure, and, when unable to do so, made a guess. These observations indicate that providing 'help' options in the case of a CA may not prove as helpful as for a GUI. While employing fallback responses to help users discover possible workarounds has been

suggested in prior work [54], users in our study bypassed the fallback mechanisms and simply tried to guess the response options instead. We therefore recommend helping users proactively rather than retroactively.

A questionnaire designed to elicit discrete responses might incorporate the expected response format into each question as suffix or prefix, for example. In the case of our discrete CA design, for instance, each question might be phrased as either “*On a scale of 0-5, how often have you felt active and vigorous?*” or “*How often have you felt active and vigorous, on a scale of 0-5?*”.

5.3.4 Exploit CAs’ interactivity to elicit richer and more complete responses. We found that users’ responses to the open-ended CA design did not always easily map to the standardized scale – potentially representing a significant challenge for the re-appropriation of standardized questionnaires to open-ended response formats. One way of mitigating this challenge may be to design conversations that allow *follow-up* intents as means of complementing users’ responses. If a user responds to a question by commenting “*half of the time*”, “*not as many days as I wanted*” or “*I tried to do things that interest me*”, for example, a *follow up* intent might ask “*Ok, so how would you rate your experience on a scale of 0-5?*” in order to confirm the user’s intended meaning. Additionally, aligning questionnaire responses with a *Yes/No* hierarchy may help to further avoid errors and render CAs more habitable. For instance, in the case of the WHO-5 questionnaire, instead of asking “*How often have you felt active and vigorous?*”, a CA might ask “*Have you felt active and vigorous?*”. A “*Yes*” response to the question can then be followed up by asking “*How often?*”, with either discrete or open-ended responses possible. A “*No*” response may be registered as ‘*at no time*’, ‘*0*’ or likewise further explored in greater detail. In this way, a CA can be designed to gather discrete and open-ended responses which, in combination, provide a more complete picture of users’ experience, spanning multiple levels of complexity and detail – as a conversation between two human actors in pursuit of mutual understanding might unfold.

6 LIMITATIONS

This study was conducted in a lab setting, meaning that participants interacted with the CA only once, and thereby limiting the extent to which our results may generalize to the context of users’ daily lives. The participant sample was also limited in terms of age and occupation. Future work should therefore consider additional demographic groups and conduct longitudinal evaluations in home settings where such devices are predominantly designed for use. It should also be noted that *Sofia* was designed to support the self-report of wellbeing as defined by the WHO-5 questionnaire. While the design implications from this study may therefore have wider implications for the self-report of wellbeing via CAs in general, their applicability to other health and wellbeing questionnaires may require further evaluation.

7 CONCLUSION

This study was undertaken with the aim of generating a richer understanding of the relationship between the design of CAs for the self-report of health and wellbeing, and users’ reporting practices and experiences. Findings reveal a strong correlation between

responses made through CA and on paper, suggesting that CAs can serve as a viable means of administering standardized wellbeing scales and collecting self-reported data. Although a CA designed to elicit ‘discrete’ responses was rated as significantly more habitable – i.e., yielding a good match between the user’s cognitive model of the system and the actual system – than an open-ended CA design, no significant differences were identified between both conditions in terms of system response accuracy, likability, cognitive demand, annoyance, or speed. Based on our empirical findings, we recommend that the design of CA applications for the self-report of health and wellbeing take into consideration users’ needs for conversational cues to support engagement, simplify questionnaire presentation, incorporate expected response formats into each question in order to mitigate guessing behaviors, and exploit the interactive potential of CAs to elicit more complete responses.

ACKNOWLEDGMENTS

We thank our participants for their participation in this study. This project is supported by the Novo Nordisk Foundation, Grant Number NNF16OC0022038, and the Copenhagen Center for Health Technology (CACHET).

REFERENCES

- [1] Matthew P Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI’14 Extended Abstracts on Human Factors in Computing Systems*. 749–760.
- [2] Ivano Azzini, Daniele Falavigna, Toni Giorgino, Roberto Gretter, Silvana Quaglini, Carla Rognoni, and Mario Stefanelli. 2003. Automated spoken dialog system for home care and data acquisition from chronic patients. *Studies in health technology and informatics* 95 (2003), 146–151.
- [3] Per Bech, Lis Raabaek Olsen, Mette Kjoller, and Niels Kristian Rasmussen. 2003. Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36 Mental Health subscale and the WHO-Five well-being scale. *International journal of methods in psychiatric research* 12, 2 (2003), 85–91.
- [4] 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.
- [5] Ava Mutchler Bret Kinsella. 2019. Voice assistant consumer adoption in healthcare. Retrieved Jan 30, 2020 from <https://voicebot.ai/voice-assistant-consumer-adoption-report-for-healthcare-2019/>
- [6] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. 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/10.1145/3313831.3376789>
- [7] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies* 139 (2020), 102410.
- [8] Yukina Chen. 2017. *The Effects of Question Customization on the Quality of an Open-Ended Question*. Nebraska Department of Education, Data, Research, and Evaluation.
- [9] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) (DIS ’19). Association for Computing Machinery, New York, NY, USA, 1557–1569. <https://doi.org/10.1145/3322276.3322332>
- [10] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. “Nobody Speaks That Fast!” An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments. 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/10.1145/3313831.3376569>
- [11] Shiu-Wah Chu, Ian O’Neill, Philip Hanna, and Michael McTear. 2005. An approach to multi-strategy dialogue management. In *Ninth European Conference on Speech Communication and Technology*.

- [12] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. *arXiv preprint arXiv:1810.06828* (2018).
- [13] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 475.
- [14] Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing* 7, 1 (2015), 74–93.
- [15] Eric Corbett and Astrid Weber. 2016. What can I say? addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 72–82.
- [16] David DeVault, Ron Artstein, Grace Bern, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.
- [17] Kevin Doherty, Andreas Balaskas, and Gavin Doherty. 2020. The Design of Ecological Momentary Assessment Technologies. *Interacting with Computers* (2020).
- [18] Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive Synthetic Speech: Voice Perception and User Behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) (CUI '20). Association for Computing Machinery, New York, NY, USA, Article 6, 9 pages. <https://doi.org/10.1145/3405755.3406120>
- [19] Hannah Gaffney, Warren Mansell, and Sara Tai. 2019. Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR mental health* 6, 10 (2019), e14166.
- [20] Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting Usability and User Experience Information from Online User Reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) (CHI '13). Association for Computing Machinery, New York, NY, USA, 2089–2098. <https://doi.org/10.1145/2470654.2481286>
- [21] Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (2000), 287–303.
- [22] Kate S Hone and Robert Graham. 2001. Subjective assessment of speech-system interface usability. In *Seventh European Conference on Speech Communication and Technology*.
- [23] Infermedica. [n.d.]. Symptom Checker. <https://www.amazon.com/Allianz-Technology-Symptom-Checker/dp/B07CS6T91V> Retrieved September 17, 2020 from <https://www.amazon.com/dp/B07CS6T91V>.
- [24] Dai Applications & Insights. [n.d.]. Wellness Guru. Retrieved September 17, 2020 from <https://www.amazon.com/dp/B075TXSNL3>.
- [25] Junhan Kim, Sun Park, Lionel Robert, et al. 2019. Conversational Agents for Health and Wellbeing: Review and Future Agendas. (2019).
- [26] A Baki Kocabali, Liliana Laranjo, and Enrico Coiera. 2018. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference* 32, 1–12.
- [27] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. 2019. The personalization of conversational agents in health care: Systematic review. *Journal of medical Internet research* 21, 11 (2019), e15360.
- [28] A Baki Kocaballi, Juan C Quiroz, Liliana Laranjo, Dana Rezazadegan, Rafal Kocielnik, Leigh Clark, Q Vera Liao, Sun Young Park, Robert J Moore, and Adam Miner. 2020. Conversational Agents for Health and Wellbeing. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
- [29] Rafal Kocielnik, Daniel Avrami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) (DIS '18). Association for Computing Machinery, New York, NY, USA, 881–894. <https://doi.org/10.1145/3196709.3196784>
- [30] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
- [31] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.
- [32] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.
- [33] Esther Levin and Alex Levin. 2006. Evaluation of spoken dialogue technology for real-time health data collection. *Journal of medical Internet research* 8, 4 (2006), e30.
- [34] Jorge Martin. [n.d.]. Depression Test. Retrieved September 17, 2020 from <https://www.amazon.com/dp/B01N11FAK9>.
- [35] M McTear, Z Callejas, and D Griol. 2016. The Conversational Interface: Talking to Smart Devices: Springer International Publishing. Doi: <https://doi.org/10.1007/978-3-319-32967-3> (2016).
- [36] Roger K Moore. 2013. Spoken language processing: where do we go from here? In *Your Virtual Butler*. Springer, 119–133.
- [37] Nasim Motalebi, Eugene Cho, S Shyam Sundar, and Saeed Abdullah. 2019. Can Alexa be your Therapist? How Back-Channeling Transforms Smart-Speakers to be Active Listeners. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 309–313.
- [38] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We Need to Talk: HCI and the Delicate Topic of Spoken Language Interaction. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) (CHI EA '13). Association for Computing Machinery, New York, NY, USA, 2459–2464. <https://doi.org/10.1145/2468356.2468803>
- [39] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2459–2464.
- [40] Cosmin Munteanu, Gerald Penn, Ron Baecker, and Yuecheng Zhang. 2006. Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. In *Proceedings of the 8th international conference on Multimodal interfaces*. 39–42.
- [41] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.
- [42] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 6.
- [43] Nadim Nachar et al. 2008. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology* 4, 1 (2008), 13–20.
- [44] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research* 19, 5 (2017), e151.
- [45] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 913–921.
- [46] Edison Research. 2020. The smart audio report winter 2019 from npr and edison research. Retrieved Jan 30, 2020 from <https://www.edisonresearch.com/the-smart-audio-report-winter-2019-from-npr-and-edison-research/>
- [47] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.
- [48] Ben Shneiderman. 2000. The Limits of Speech Recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. <https://doi.org/10.1145/348941.348990>
- [49] Arthur A Stone and Saul Shiffman. 2002. Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine* 24, 3 (2002), 236–243.
- [50] Bernhard Suhm. 2003. Towards best practices for speech user interface design. In *Eighth European Conference on Speech Communication and Technology*.
- [51] Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. 2015. The WHO-5 Well-Being Index: a systematic review of the literature. *Psychotherapy and psychosomatics* 84, 3 (2015), 167–176.
- [52] John Torous, Patrick Staples, Meghan Shanahan, Charlie Lin, Pamela Peck, Matcheri Keshavan, and Jukka-Pekka Onnela. 2015. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR mental health* 2, 1 (2015), e8.
- [53] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.
- [54] Zhuxiaona Wei and James A Landay. 2018. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.
- [55] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2019. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys. *arXiv* (2019), arXiv–1905.