HarborBot: A Chatbot for Social Needs Screening

Rafal Kocielnik[1], Elena Agapie[1], Alexander Argyle[1], Dennis T. Hsieh[2], Kabir Yadav[2], Breena Taira[2], Gary Hsieh[1]
[1]University of Washington, WA; [2]UCLA Medical Center, CA

Abstract

Accessing patients' social needs is a critical challenge at emergency departments (EDs). However, most EDs do not have extra staff to administer screeners, and without personnel administration, response rates are low especially for low health literacy patients. To facilitate engagement with such low health literacy patients, we designed a chatbot - HarborBot for social needs screening. Through a study with 30 participants, where participants took a social needs screener both via a traditional survey platform and HarborBot, we found that the two platforms resulted in comparable data (equivalent in 87% of the responses). We also found that while the high health literate participants preferred the traditional survey platform because of efficiency (allowing participants to proceed at their own pace), the low health literate participants preferred HarborBot as it was more engaging, personal, and more understandable. We conclude with a discussion on the design implications for chatbots for social needs screening.

Introduction

Although designed for medical emergencies, emergency departments (EDs) have become a common place where patients seek help for various problems. EDs thus care for patients who present for not only heart attacks and strokes, but also a variety of social ills, such as homelessness, poverty, and hunger[1]. These social concerns disproportionately affect the poor and limit their access to healthcare. Left unaddressed, social determinants drive up cost and utilization, and are a fundamental source of persistent health disparities.

However, despite the growing body of evidence that addressing social determinants improves health outcomes[1] , assessing patients' social needs remains a critical challenge. Most EDs currently do not screen for social needs and do not have extra staff to administer screeners without interrupting clinical workflow. Without personnel administration (such as research assistant, nurse, etc.), response rates for both paper and electronic surveys are low[2]. This is compounded by the fact that only 12% of Americans have proficient health literacy[3].

Chatbots offer multiple potential benefits for social needs screening. Chatbots are systems designed to engage with users through natural language, mimicking a human-to-human interaction. Popular examples of chatbots include Apple's Siri, Google's Now, and Microsoft's Cortana. Extended to the context of social needs assessment chatbot can support self-administering of social needs screeners to minimize personnel cost. In contrast to current form-based survey, a conversational approach would be more "chat" like, potentially offering a sense of familiarity similar to mobile text messaging. By creating a sense of interacting with another person, the chatbots may also increase participation engagement. Furthermore, offering text-to-speech output can also facilitate comprehension.

In this work, we present HarborBot, a chatbot for administering social needs screeners. HarborBot's persona is designed to be professional, trustworthy, and supportive. It uses a chat-like interface to ask the social needs screener, and allows survey takers to ask for clarifications as if they are interacting with a survey administrator. To study its use, we recruited 30 participants, which includes both high and low health literacy individuals, with both ED and non-ED patients. Through a within-subjects study design, participants were asked to take the survey twice, once via HarborBot, and once via an existing survey platform, SurveyGizmo (order randomized).

Our finding suggests that there is a clear divide in preference between the high health literate (HL) and low health literate (LL) participants. Almost all HL participants preferred taking the survey via SurveyGizmo, whereas the LL participants preferred HarborBot. Post-study interviews reveal that HL individuals consider conversational approach to be too slow without allowing them to take the survey at their own pace. On the other hand, LH participants appreciated the audio output of the HarborBot, and thought the conversational design was much more engaging and personal.

Our work offers numerous insights on the role of conversation-based approach for surveys, and advance our understanding of how to more effectively screen for social needs for both high and low health literacy patients.

Related Work

With the growing interests in clinical screening, research has examined the use of technology-based solutions to support the self-administering of surveys, or Computer-Assisted Self-Interviewing (CASI)[2]. CASI include the use of online survey platforms, mobile apps, and electronic kiosks. These self-administered solutions may help maximize scalability and speed of data collection while reducing cost. Compared to face-to-face interviewing, self-administered solutions can also reduce social desirability bias, and limit the under-response in sensitive issues[2]. Furthermore technology-based solutions led to fewer item missing responses, with the same response rates, compared to pen and paper self-administered questionnaires (PAPI)[4].

A variation of CASI is Audio Computer Assisted Self-Interviewing (ACASI). The ACASI system allows users to listen to pre-recorded or text-to-speech audios of questions as if they were being asked by experimenters. This is especially valuable for supporting the understandability among low literacy participants[5]. The use of ACASI also reduces social desirability bias, making it an effective format to collect sensitive information[6]. This is valuable when assessing social needs in the diverse ED population. One drawback is that ACASI systems take longer to interact with[7].

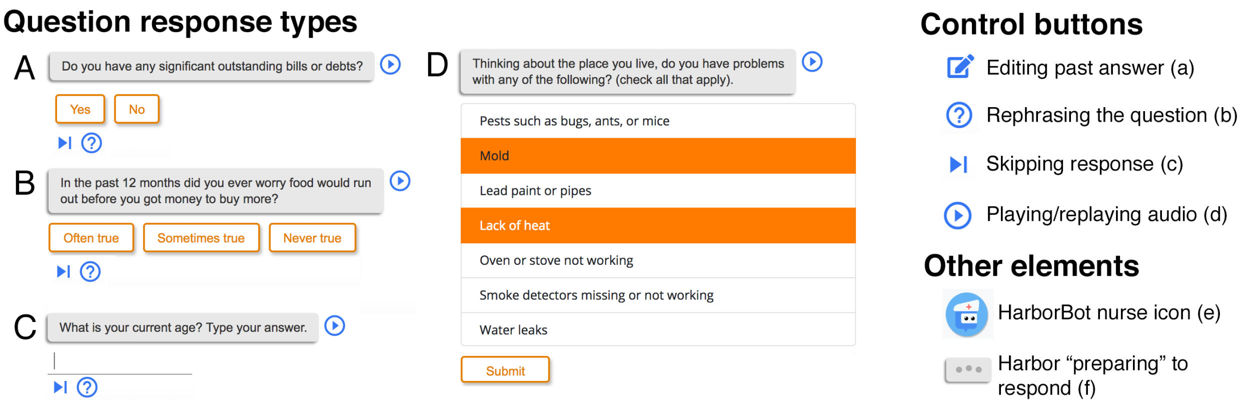
Despite these advantages of existing technology-based solutions, face-to-face is still better when it comes to response rates (in one study, 92.8% face-to-face response rate compared to 52.2% web-survey response rate)[8]. These differences, also in ED context[9], have been linked to the motivating impact of interpersonal interactions, but reproducing such effects via technology is still a challenge. Additionally, non face-to-face surveys often bias against nonwhites[10], low income patients, homeless or those that are disenfranchised with mental health and/or substance use[9].

The Potential Role of Chatbots in Survey Screening. The idea of using an interactive system to collect and deliver health-related information has been studied primarily via so called embodied conversational agents (ECA)[11]. ECAs allow people to interact with physical or graphical agents that embody a person in appearance, behavior, and dialect. Mimicking a human-like interaction, ECAs are able to improve engagement and trustworthiness[12]. They are also perceived by patients with varying literacy levels as acceptable and easy to use[13], However, their potential for screening is less well studied. Only a few research systems have been prototyped to explore patients interviewing[14], and the actual use of these systems is quite limited, as the cost of development these human-like embodiments is high.

One potential lighter-weight alternative to ECA are chatbots which are composed of six key features[15]. First is the concept of thread as app, where the app-centric homescreen is replaced by a threaded conversation with streams of messages and notifications. Much like the interaction one has with another through chat messages. Second is history awareness, where the bot keeps a log of past interactions with users in the thread. Third is an enhanced user interface (UI), where the interaction need not be limited to plain text, but can also include images, audio output, structured messages. Fourth is limited natural language processing (NLP), which is to prevent breakdowns due to technology limitations. Still, chatbots try to mimic conversation with a human partner, e.g., by using ”is typing” indicators, dynamic utterances and persona-driven emotional responses. Fifth is message self-consistency, where the intent of each message is clear and stands on its own. Finally, use of guided conversations to prevent users from getting lost.

Like ECAs, chatbots have multiple features that can be ideal for a low literacy population. For example, the audio output can facilitate understandability; the conversation-like interactions can foster a sense of interacting with another person, making the interaction more personal and engaging. But chatbots also differ from ECA in a couple of critical ways. One is that chatbots do not require a graphical or physical embodiment of the agent. This minimizes the cost of development and offers a more scalable solution. People can interact with chatbots even through their mobile devices. Another key difference is that chatbot mimic text messaging. Demographic surveys have shown that people with lower SES (more likely to have lower health literacy and higher social needs) are increasingly reliant on mobile text messages to communicate[16]. The chat-like interface may be more welcoming and intuitive for those users.

Chatbot based surveys were shown to increase user engagement and produce higher quality responses[17]. Such effects, however, have only been demonstrated with low-stakes demographics and marketing online surveys, and with general population. It is unclear if and how such benefits could be translated to clinical setting and low literacy population. Furthermore, low literacy in ER poses unique challenges in term of understandability and comfort with sharing information. Finally, prior work has provided very little in terms of linking the particular chatbot design aspects to their effect on the users, which limits its value for informing future design decisions. Our work aims to fill these gaps.



Do you have any significant outstanding bills or debts?

In the past 12 months did you ever worry food would run out before you got money to buy more?

What is your current age? Type your answer.

Thinking about the place you live, do you have problems with any of the following? (check all that apply).

- Pests such as bugs, ants, or mice
- Mold
- Lead paint or pipes
- Lack of heat
- Oven or stove not working
- Smoke detectors missing or not working
- Water leaks

Figure 1: HarborBot GUI elements. On the left “Question response types” showing different types of responses users available. On the right “Control buttons” show the 4 controls associated with each question. “Other elements” show HarborBot icon and an ellipsis icon HarborBot used for mimicking writing by a person in chat interaction.

System: HarborBot

We designed and implemented a custom chatbot called *HarborBot* to test a conversational approach to surveys. HarborBot interacted with users through chat and voice. It communicates via chat messages, that it can also read out, as if it is speaking. Users interacted with the system primarily through buttons (for structured responses) and text (for text-based questions). HarborBot is implemented as a webapp and we had our participants interact with it on tablets.

Design Process. To create HarborBot we followed an iterative design process in which a team of 2 senior HCI researchers and 6 design students followed three general design phases: 1) Requirements gathering - the team consulted 3 ED practitioners, who are also co-authors on the paper, and existing literature related to patient experience in ER[9], 2) Design exploration - the team explored various low-fidelity prototypes and gathered feedback on them from ED practitioners and via small scale usability tests, 3) Refinement - the most promising prototype was developed further and refined with positive elements from other prototypes. This process resulted in the final HarborBot system.

User Interface. We used BotUI - a Javascript framework to build conversational UIs[1]. Messages to and from HarborBot appear in standard elliptic chat bubbles with users messages distinguished from the bot’s by different colors. Prior to the appearance of messages from HarborBot, animated ellipses are shown in the chat bubble with a delay to denote that the bot is typing (Fig 1 f), akin to that of iMessage or similar interfaces.

BotUI allows for the creation of different question types, which we used to cover the types of questions asked by the screener. Each message from BotUI would be one of these types: skip, yes/no, input, options, or many options. Skip (see Fig 1 c) was a unique type that would move onto the next message of HarbotBot’s script without user responding. Yes/no (Fig 1 A), and options (Fig 1 B) use the standard buttons offered by BotUI. For questions involving multiple possible answers (i.e., checkboxes), we used a vertically stacked list of options that allowed users to choose multiple options before submitting (Fig 1 D). Free response (Fig 1 C) required a text field and device’s keyboard to respond.

Persona. There were several personality considerations we made in the design of HarborBot. Most importantly we emphasized striking a balance between a serious and friendly tone. If HarborBot was too friendly, users may feel the conversation is not being taken seriously. If too serious, users will lose the feeling of comfort we want them to have in answering personal questions. We also intentionally avoided any use of humor seeing as it would be inappropriate in the context. In addition, given the stressful nature of patients’ experiences in EDs, we sought to make HarborBot empathetic. However, we tried to do so without pitying the user or being condescending in any way. To accomplish this, HarborBot used occasional confirmatory phrases, such as: *“Okay, I’m getting a better idea of where you are at.”*, *“Got it”*, and assurances, such as: *“The next questions are about your personal safety and may be tough to answer.”*

The voice of HarborBot was an important part of the interaction with the user. By default, HarborBot used a female

[1] BotUI - https://botui.org/

voice taken from the Microsoft's Bing Voices[2]. Users could adjust the volume of the voice or mute it entirely for privacy reasons or personal preference.

Dialogue-Based Interactions. We employed thread as app principle to support survey-taking. Survey questions and user replies were presented as streams of messages in threaded conversation akin to chat messaging. We allowed users to skip and ask for clarification of the questions. Each question posed by HarborBot offered a couple of utility options using round buttons next to the answer area. The Skip button (Fig 1 c) allowed the users to skip questions that felt uncomfortable or they couldn't answer. After skipping, HarborBot would proceed on with the script. We supported rephrasing the question to offer its simplified version for low literacy individuals (Fig 1 b). We used Readable to make sure our rephrased versions were at most at a fifth grade reading level. Selecting this option is akin to saying *"I didn't quite understand that."*, which would make HarborBot re-ask the question using the simplified phrasing.

Additionally, we implemented an edit button (Fig 1 a) next to each past answer in case user needed to change it. Selecting this option would make HarborBot ask the corresponding question again and allow the user to provide a new answer. This is akin to users saying to the bot "Can I change my response to that question?"

Occasionally, HarborBot would respond with conversational remarks. These utterances were essential to developing Harbor's personality, and engaging users in a conversation. Some of these interactions are dynamic based on a rule-based approach. For instance, if a user indicated they did not have a steady place to live, HarborBot would not ask the remaining housing questions. If the user response indicated a negative social situation, HarborBot would acknowledge it with a sympathetic affirmation, such as *"That must be stressful, I'm sorry to hear that."*

Method: Study

We conducted a within-subjects study with 30 participants to compare the experience of answering a social needs survey using two different platforms: HarborBot (Chatbot) and a more traditional interface for taking surveys - Surveygizmo (Survey). We recruited participants with high and low health literacy at two study sites. We expected the Chatbot interface will be 1) more engaging, 2) more understandable, and 3) more comfortable to share information with, while 4) preserving response quality. We also expected these effects to be pronounced with low literacy users.

Study Procedure. Users interacted with both survey interfaces using a tablet's web browser. After interacting with one interface, participants reported their perceptions and experience. They then repeated the same procedure for the second interface. We randomized the order of interaction. After completing both, we conducted an interview.

Social Needs Survey. In both platforms users answered the social needs survey developed by the Los Angeles County Health Agency (LACHA)[18]. This survey comes from over two years of work from the committee on the Social and Behavioral Determinants of Health, with members from the Departments of Health Services, Mental Health, and Public Health. The survey asks 36 questions related to demographics, financial situation, employment, education, housing, food, and utilities as well as questions related to physical safety, access to care, and legal needs. A number of questions can be considered sensitive, such as: *"Have you ever been pressured or forced to have sex?"*, *"Are you scared of being hurt by your house?"*, *"Did you skip medications in the last year to save money?"*

Measures. Participants evaluated both survey platforms in terms of workload (NASA TLX survey[19]), engagement in the task (questions adapted from O'Brian's engagement survey[20], e.g., *"I was really drawn into answering questions."*, *"This experience of answering questions was fun."*, *"I was absorbed in answering questions."*), understandability of content (*"I understood the questions that were asked of me."*), and willingness to share information (*"I was comfortable answering the questions."*). These measures have been commonly used in prior studies of chatbots[13].

Participants health literacy was measured using Rapid Estimate of Adult Health Literacy (REALM)[21] which assesses participant's ability to read health materials and instructions, at a comprehension level of high school or lower[21]. We used the Newest Vital Sign (NVS) health literacy scale[22] to assess likelihood of limited health literacy based on numeracy, prose and document literacy measures. NVS was used for fast recruitment (under 3 minutes).

During interview we asked about preferences for the two survey platform, the specific features of the platforms, participants' comfort in sharing information in each platform, and perceptions of the personality of the chatbot.

[2] https://docs.microsoft.com/en-us/azure/cognitive-services/speech/api-reference-rest/bingvoiceoutput

Recruitment. We recruited participants from two study sites. In the Seattle metropolitan area participants were recruited through Craigslist, flyers at local community centers, and the Institute of Translational Health Sciences's research subjects' pool. In the Los Angeles County participants were recruited from a large county safety net hospital (Harbor-UCLA), by two of the authors who are physicians there. ED visitors were handed out a flier at discharge, or while waiting. Participants were 18 or older and had a conversational level of English proficiency. The study was approved by the IRB at both sites.

Participants. 30 participants were recruited (17 males, 10 females, 3 declined to answer) ranging from 23 to 65 years of age (M=39.63, SD=12.91). They reported completing 13.15 (SD=3.73) years of education on average. 22 reported English and 4 Spanish as their primary language. One person reported bilingual fluency and 3 people declined to answer. Finally we had a diverse ethnic backgrounds: Hispanic or Latino: 9, Black or African American: 8, White: 6, Multi race: 2, and 4 reported other ethnicity or declined to answer.

11 participants were assessed as low, and 19 as high health literacy. Participants were considered low literacy if they scored at a seventh to eighth grade level or below, on the REALM scale, or got a score that suggests high likelihood (50% or more) of limited literacy on the NVS scale.

Analysis. Quantitative data comprised user responses for Chatbot and Online Survey interface, Chatbot interaction logs, and post-interaction survey responses for each platform. These were matched by participants' unique id and the analysis focused on descriptive statistics of user interactions, especially with Chatbot, and on comparison of answer equivalence for the two platforms. Differences in survey responses were assessed using paired t-tests and interactions between interface type and participant's health literacy levels were explored using linear mixed effects models.

The interviews took between 7 and 25 minutes (M=17.56, SD=9.21), conducted by three and analyzed by four of the authors. Each researcher wrote a detailed summary of interviews they had not conducted, including quotes. We then developed a codebook following a *top-down* and *bottom-up* approaches. Initial codes for the *top-down* pass were informed by the interview questions (why participants liked or disliked each survey platform, attitudes towards sharing information, perceptions of Chatbot). The interview structure itself was informed by the literature. We then refined the codes based on themes that emerged from the data in a *bottom-up* fashion. Each interview summary was coded by a researcher on the team (who had not conducted the interview, or written the summary). The coded interview summaries were used to identify themes. Three of the authors discussed the overall themes until consensus was reached. Researchers consulted with the audio and transcriptions of the interviews to ensure validity of the coding.

Quantitative Results

Preferences. Low health literacy (*LL*) participants preferred using Chatbot over the Survey with 8 out of 11 expressing such preference. At the same time, 17 out of 19 high literacy (*HL*) participants preferred Survey. This difference was statistically significant ($\chi 2$ (2, N = 30) = 12.5, p <.001).

Time to Completion. Participants had to respond to 36 questions in the social needs survey, but they could also skip answers. They spent significantly (t(27)=2.23, p<0.05) more time answering questions via Chatbot (M=9:26 min; SD=3:14 min) than via Survey (M=6:48 min; SD=6:28 min). We found no significant difference between answering time (avg. of both interfaces) for LL (M=9.43 min, SD=3.23) and HL participants (M=7.36 min, SD=4.20). We also found no significant interaction between the interface and literacy level on time.

Equivalence of Responses. An important question is whether the two interfaces result in the same data quality. We explore two measures: *per-item response rates* and *data equivalence*. On average participants provided almost identical number of answers via the two interfaces: 32.93 (SD=3.48) questions answered with Chatbot and 33.00 (SD=2.95) with Survey. This suggests *comparable response rates*. In terms of data equivalence 87.0% (SD=11.6%) of the responses per user were the same across the two interface versions.

Reasons for Response Discrepancies. We found that skipping an answer in one interface, but not the other was the primary cause of answer discrepancy (48% of mismatches). There was, however, no significant difference between the two platforms in skipping behaviors. 25% of mismatches was a result of skipping a question in Chatbot only and another 23% due to the oppsite. Furthermore, the order in which users encountered the interfaces had no significant impact on skip rates: 8.0% (SD=9.3%) when answering the survey the first time, and 7.8% (SD=8.2%) when answering the survey the second time. Hence the platforms are not different in this respect. One interesting finding from our explorations is that there seems to be an anchoring effect with users skipping more often when starting the study with Chatbot, for their responses to both platforms: Chatbot (M=9.8%, SD=29.7%) and Survey (M=9.8%, SD=27.2%) than when starting with Survey: Chatbot (M=5.2%, SD=22.3%) and Survey (M=4.9%, SD=21.6%). This is most likely due to the skip option being more explicit in the Chatbot and users wanting to be consistent in their answers.

Manual examination of the remaining mismatches revealed varied and non-systematic reasons for discrepancies such as: low equivalence only in the very first introductory question (53.3%), direct contradiction (e.g., user answered "Yes" in one interface and "No" in the other); similar, but not the exact same answers (e.g., answer: "Yes, help finding work" vs. "Yes, help keeping work"), ticking an additional option in a multi-choice answer (e.g., "Unemployed - looking for work" vs. "Unemployed - looking for work, Disabled") and a possible misinterpretation of the question (e.g., when asked for income per month, user typed "2000" in one interface and "24,000" in the other).

Workload (NASA TLX). Analysis of the NASA TLX survey responses revealed a difference in task load index (avg. of all items denoting workload, α: 0.83) between Chatbot and Survey. Participants reported a higher workload when using Chatbot (M=2.460, SD=1.241), compared to Survey (M=2.167, SD=1.284; t(27)=-2.020, p=0.05). Given the scale from 1–lowest to 7–highest, this still represents a low perceived workload. We also found a main effect of literacy level: there was a higher perception of workload across both platforms by the LL participants (M=2.955, SD=1.335) than the HL ones (M=1.921, SD=0.948; t(27)=2.439, p<0.05). The interaction effect was not significant.

Engagement, Understandability, and Comfort with Sharing Information. Analysis of the engagement index (average of O'Brian's engagement questions, α: 0.82), revealed a higher reported engagement for LL participants (M=3.920, SD=0.502) than HL ones (M=3.469, SD=0.402), (t(27)=2.672, p<0.05). We also found a weakly significant interaction between interface and literacy with LL participants being more engaged with the Chatbot than HL ones, but less engaged with the Survey (Chatbot*Low, β=0.485, SE=0.262, p=0.064). This represents a half a point increase on a 5-point likert scale for engagement. Trends in the same direction, but no significant differences were found for *understandably* and *comfort with sharing information*.

Qualitative Results

In this section, following mixed-methods approach, we complement and expand on the quantitative findings. Participants varied not only in their preferences for Chatbot or Survey, but also in the particular aspects they liked about each, as well as in which design aspects were instrumental in creating particular perceptions and experiences. Participants valued the engaging conversational aspects of the Chatbot. Especially LL participants found the conversational interface more caring in the context of a sensitive topic. In contrast, HL valued the efficiency of the SurveyGizmo interface and felt slowed down by the Chatbot. Some participants found the Chatbot more robotic, disingenuous or pushy at times, but these seem to result from the particular way in which HarborBot implemented conversation.

Strengths of Our Conversational Approach

Engaging. Most participants found the conversational features of the chat more engaging than the Survey, regardless of the health literacy level. Participants felt like they were having a conversation with a person when using the Chatbot. More than half the participants attributed such perception to the use of *voice*: *"she was reading the questions and I can answer it ... seemed like a conversation ... like someone was talking to me and it gave me the opportunity to answer back and then they answered back"* (H59). Other participants felt the *ellipses* made it feel like having a chat with someone (H76, L77), and even referred to the the Chatbot as *"she"* (8 participants). Some participants valued that the Chatbot felt like a person: *"I liked... how it talked to you, reads you the questions ... it spoke directly at me"* (L60), *"I thought it was someone asking me those questions"* (L72).

Aside from the voice and ellipses, the *conversational utterances* also contributed to the perception of interacting with a person (L75, L58, L72, H32, L60, L36, H41, H59). One participant found them motivating: *"Saying 'you got it.' It's giving you motivation ... nice to hear that once in a while"* (H73). Another felt like the conversation was adapting to the answers to be more relevant: *"seem like they tried to give you a little positiveness based on your answer"* (H59).

Caring. Participants perceived the Chatbot as caring, particularly in the LL group. These participants had a generally positive attitude towards the social needs survey questions (L51, L55, L58, H73) and this topic resonated with their personal experiences *"It felt like it was telling me about my life. That was really amazing, like woow"* (L71). Therefore, some of the perceptions of the Chatbot might have been accentuated by the positive perception towards the survey topic. Many participants described the personality of the Chatbot using terms such as: caring, kind, patient, helpful, calm, familiar, or concerned (H35, H41, L52, L55, L57, L61, L77). Participant also reported the *voice* of the Chatbot was aligned with this caring personality: it was soothing (H57), had cadence (H32), helped a nervous participant feel more comfortable (L55) and was *"nice and sweet made me feel relaxed"* (L77).

The Chatbot was designed to provide *supportive utterances* in response to some of the participants answers. Many participants liked these utterances (L60, L36, H32, H41, H58, H59). One participant though the utterances made him feel *"comfortable to answer the questions"* (L61), and that they provided a positive reinforcement to keep on answering (H59). Participants perceived Chatbot utterances such as *"I am sorry to hear this"* as the Chatbot *"trying to be understanding"* (H59). Some found these utterances to be very applicable to the conversation context. For example L61 considered the Chatbot response: *"That must be stressful"* to be a reaction to the information she shared: *"she probably said that because of my financial situation"* (L61), which she felt would be calming for people *"to not be stressed, I would think it would be helpful"* (L61). Other participants felt the supportive utterances gave them confidence: *"nice lady giving me confidence ... with good tone of voice"* (L75).

Understandable. Several LL participants (5 of 11) reported having trouble with reading and understanding the written questions in the Survey. They liked using the Chatbot because it facilitated their understanding, which they attributed to the audio feature: *"When I hear it I have a better understanding of the question"* (L61) or that *"just hearing it I could ... relate better to the question"* (L53). Some participants reported using the feature that replayed audio, to better understand a particular item (L51, L58, L61, L73). This was especially useful when they missed some words or did not fully comprehend some of the contents at first: *"I didn't get it at first, so I wanted to go back and listen to it again before answering"* (L58). Several also mentioned that they would have liked it if the answers were spoken via audio as well, to make them more understandable (L61, L54).

Accessible. Some participants had particular needs that the Chatbot was able to satisfy much better than the Survey. One participant who reported vision problems, preferred having questions read to them: *"If it is too small I can't see it so I prefer to have the questions read to me anyways"* (H73). Another participant reported feeling very comfortable with the Chatbot because she was regularly experiencing panic attacks and considered ED stressful: *"I was thinking I was texting somebody ... that made me forget where I was at ... it was like texting my sister my mom and waiting for them to respond back. And that made me feel patient"* (L77). In contrast, she found it particularly difficult to take the Survey: *"by myself ... it felt awkward and alone"* (L77).

Weaknesses of Our Conversational Approach

Inefficient. HL participants cared about efficiency, primarily reflected in the speed of completing the survey. The majority of HL participants (17 out of 19) preferred to use the Survey because of that. Several mentioned that the traditional interface enabled them to be faster than the Chatbot (H21, H22, H24, H59), or to go at their own pace (H36). Participants attributed being slowed down to various conversational features of the Chatbot. Some felt the Chatbot was slower because they needed to wait for the ellipses before a new question would appear (H35). They were also able to read faster than the questions were read by the bot: *"when she was talking at me. I felt like I was going at a slower pace"* (H23). Also not having to engage with additional conversational utterances was seen as more efficient (H35, H56). The audio feature was perceived as interfering with reading and thinking (H23, H40, H70, H21, L71). One LL participant preferred the Survey because they could concentrate more: *"to read is better ... Because that way I could like concentrate more and think about more and you know ... I could read my letters more and makes it better for me."* (L58).

Pushy. Somewhat surprisingly, a few participants perceived Chatbot as being pushy, based on the tone and the speed at which questions were asked. Some participants felt the questions asked were very direct (H57, L72, H52). L72 felt like he was answering questions to a teacher, and had to provide correct answers. H57 and L72 thought there could be more utterances to help prepare the survey taker for some very sensitive questions in the survey. H57 also felt that some of the questions were trying to repeatedly get information that he had already declined to provide: *"if I say none of the above ... don't be pushy"* (H57). Others also felt rushed in providing the answers to the Chatbot. For example, the use of ellipses, and the short delay between its messages made it feel like the Chatbot was moving faster than the participants were comfortable with (H23, H63). Participant H63 felt like the questions kept coming and he had no control over when they would be read.

Robotic and Disingenuous Voice. Some participants, primarily in the HL group, perceived the Chatbot as being robotic. Some participants found the voice not sounding natural (H21, H22, H23, L58, H59, H63, H70, H76), for example sounding *“truncated .. monotone...seemed pretty artificial to me.”* (H70). Some perceived the Chatbot as disingenuous when the utterances did not meet their intended purpose (H63, H40, H23, H52): *“I feel like they were trying [to make] the software to feel sympathetic, or empathetic, that was weird”* (H63). Another participant perceived utterances as defaults: *“it felt like defaults rather than someone ’feeling for you”’* (H40). The perception of artifical responses led another participant to perceive the Chatbot as fake, and was reminded of customer support: *“kind of just programmed, recorded in, to appear to be more personal...hell there’s nobody there somewhat disingenuous ... It reminded of ... dealing with the phone company”* (H70).

Inconclusive Impact on Willingness to Disclose Information. Most participants, regardless of health literacy level, reported being comfortable sharing information asked by the survey questions. However, the human-like interactions of Chatbot did affect some participants’ willingness to disclose information, although participants reported effects in both directions. For some, if they thought they were interacting with a person, they felt more reluctant to share sensitive information, or tell the truth: *“I might be more honest if I’m reading [the question] ... if someone else ask me about them, I might lie”* (L72). Another participant showed concern about the identity of the potential conversational partner: *“it was a robot, I didn’t mind, but I think if it was a human being I would mind... and you really don’t know who’s on the other end”* (H40). In contrast, some participants were more willing to disclose because of the human-like interactions. *”If it says ’I would like to more about you’. It gives me the confidence to open up, because each question that follow sounds so interesting and it gives me the opportunity to interact with the person on the other side ... it wettens my appetite to give out more information”* (L75).

Discussion

Main Findings. In this paper, we proposed the use of a chatbot (HarborBot) for social needs screening at emergency departments and compared it to a traditional survey tool (SurveyGizmo). Based on interviews, interaction logs, and survey responses we demonstrate that the conversational approach is perceived as more engaging by all the participants, and further as more caring, understandable, and accessible among the low health literacy (LL) ones. Importantly, we also demonstrate that the conversational approach results in similar response rates and 87% equivalence in the collected data. At the same time, we found the conversational approach to be more time consuming (in line with reports from prior work on ACASI[7]) and prone to be perceived as somewhat pushy, robotic, and disingenuous which was, however, mostly the perception of participants with high health literacy (HL).

Positive Design Aspects. Numerous strengths of the conversational approach for LL population can be linked to conversational features. First, various features of the chatbot facilitate understanding. The audio output is especially valuable for participants who are less proficient readers. Second, the ability to ask the bot to rephrase the question offered a way to ask for clarification that is currently not a feature in online survey platforms. Third, chatbots can create a sense of interacting with a human. The utterances can make the survey takers feel cared for and engaged. Such positive interactions made some participants feel relaxed and even motivated to answer more questions.

Challenging Design Aspects. We were surprised that the conversational features felt pushy for some, especially HL participants. Such perception was linked to the tone of the questions and to the speed of the interaction. In terms of tone it is possible that our literal use of the wording of the survey questions was not the most appropriate for creating a conversational feel. In terms of speed of interaction, the use of voice might be a contributing factor. As reported in prior work agent asking questions via voice can create a perception of response urgency[23]. This could be improved by adding assurances like *“please take your time.”*, manipulating intonation, or making it more explicit that the ellipses represent someone is typing (rather than the system is waiting for a response). The second reason for pushy feel could be related to the fixed speed of conversation. Human-human conversation involves not only exchanging information, but also coordinating various aspects of the exchange, e.g., its speed[13]. If a participant needs more time to think, a real person, would pick it up from verbal and non-verbal cues and adjust the speed. Our HarborBot is currently incapable of making such adjustments. Such fixed speed may feel too fast or too slow for some users.

HarbotBot felt “caring” for LL and “robotic” for HL participants. This might be related to the different expectations and tolerance levels for voice quality and may be improved with use of a better quality text-to-speech service (technical challenge), human pre-recorded audio clips (which comes with limitations in flexibility), or modifications of intonation and prosody using approaches such as Speech Synthesis Markup[3]. Another way may be to generate more personalized and diverse utterances[24].

Future Design Directions. Given the division of preferences for chatbot/survey between the HL and LL groups, one possibility for a real-world use could be to have two versions of the tool and either intelligently assign or have patients pick the version they would prefer to engage with. While, long waits in healthcare setting make it less of a problem, a number of design opportunities can still be explored to make the chatbot interactions more efficient, such as simplifying the script, or providing user control over time between messages. While we focused on examining the effects of the conversational approach for a LL population, our findings suggest a potential for accessibility-focused uses of the chatbot. Participants who were hard of seeing mentioned they appreciated the audio output. Further, one participant with anxiety attacks appreciated the human-like interactions, which made them feel like chatting with a loved one, at home.

Finally, it is not clear based on our results, how the conversational approach affects people’s comfort in responding to questions, and any potential desirability biases. Prior work suggests that the self-administered screeners would reduce social desirability bias, and limit the under-response in sensitive issues[6]. This is because people will not feel like someone is monitoring or judging them. We thought the Chatbot may strike a happy medium between being perceived as human-like to enhance engagement, while not being perceived as a person for people to feel uncomfortable with disclosures. It is not clear if we were able to achieve that balance. Some participants who thought the Chatbot was human-like did not mind sharing and commented that it was more motivating, while others that thought the Chatbot was human-like were concerned with sharing. It is possible that the very initial greeting from the bot sets the tone for the rest of the interaction[25]. This requires additional research.

Limitations

Recruiting low health literacy participants with basic English proficiency is difficult. Thus our sample size for this group is limited. Still, our results appear fairly robust; our claims are supported through both quantitative and qualitative data. With our quantitative data, even with the small sample size, our primary outcome variables achieved statistical significance. Also with our qualitative results, we reached data saturation early.

We were able to study actual ED patients interacting with HarborBot, which boosts the ecological validity of our findings. Nonetheless, a number of important generalizability questions require future work. For example, would our general findings hold with a different population (e.g., non-English speakers, different cultural backgrounds), or a different set of questions? Furthermore, while we focus on examining the acceptability of chatbot for social needs screening in the ED, a critical next step is to study its feasibility and how it may be integrated into existing workflows.

Conclusion

In this paper we have proposed the use of a chatbot as a tool for social needs screening in emergency departments. We designed and built HarborBot to enable a more human-like interactions during the self-administering of surveys. Through a mixed-methods study with 30 low and high health literacy participants recruited at two different sites, we showed that compared to traditional online survey, the conversational interface offers benefits such as increased engagement, facilitating understandability and making the interaction feel more personal and caring. These benefits were especially appreciated by low health literacy participants. However, the high health literacy ones mostly preferred traditional survey as it was more efficient. Our work advances the understanding of conversational agents, and offers valuable insights on the design and potential role of chat interfaces for social needs screening.

[3] https://www.w3.org/TR/speech-synthesis11/

References

1. Patrick W Malecha, James H Williams, Nathan M Kunzler, Lewis R Goldfrank, et al. Material needs of emergency department patients: a systematic review. *Academic Emergency Medicine*, 25(3):330–359, 2018.
2. Ann Bowling. Mode of questionnaire administration can have serious effects on data quality. *Journal of public health*, 27(3):281–291, 2005.
3. Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. *National Center for Education Statistics*, 2006.
4. Anne M Johnson, Andrew J Copas, et al. Effect of computer-assisted self-interviews on reporting of sexual hiv risk behaviours in a general population sample: a methodological experiment. *Aids*, 15(1):111–115, 2001.
5. James N Gribble, Heather G Miller, Susan M Rogers, and Charles F Turner. Interview mode and measurement of sexual behaviors: Methodological issues. *Journal of Sex research*, 36(1):16–24, 1999.
6. Laura Gottlieb, Danielle Hessler, Dayna Long, Anais Amaya, and Nancy Adler. A randomized trial on screening for social determinants of health: the iscreen study. *Pediatrics*, pages peds–2014, 2014.
7. David S Metzger, Beryl Koblin, et al. Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. *American journal of epidemiology*, 152(2):99–106, 2000.
8. Dirk Heerwegh and Geert Loosveldt. Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public opinion quarterly*, 72(5):836–846, 2008.
9. Helen Chiu, Nadia Batara, Robert Stenstrom, Lianne Carley, Catherine Jones, et al. Feasibility of using emergency department patient experience surveys as a proxy for equity of care. *Patient Experience Journal*, 1(2):78–86, 2014.
10. Ingrid Llovera, Mary F Ward, James G Ryan, et al. A survey of the emergency department population and their interest in preventive health education. *Academic emergency medicine*, 10(2):155–160, 2003.
11. Ameneh Shamekhi et al. Augmenting group medical visits with conversational agents for stress management behavior change. In *International Conference on Persuasive Technology*, pages 55–67. Springer, 2017.
12. Raoul Rickenberg et al. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *Proc. of SIGCHI conference on Human Factors in Computing Systems*, pages 49–56. ACM, 2000.
13. Timothy W Bickmore, Laura M Pfeifer, Donna Byron, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *J. of health communication*, 15(S2):197–210, 2010.
14. Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, et al. Embodied conversational agent-based kiosk for automated interviewing. *J. of Management Information Systems*, 28(1):17–48, 2011.
15. Lorenz Cuno Klopfenstein et al. The rise of bots: a survey of conversational interfaces, patterns, and paradigms. In *Proc. of the 2017 Conference on Designing Interactive Systems*, pages 555–565. ACM, 2017.
16. Pew Research Center. Mobile Fact Sheet. http://www.pewinternet.org/fact-sheet/mobile/, 2018.
17. Soomin Kim et al. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *CHI 2019*, page 86. ACM, 2019.
18. Hong C Hsieh D. Liu P, Johnson S. White paper: Development of a social and behavioral determinants of health short screener. *Los Angeles County Health Agency*, 2018.
19. Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.
20. Heather L O'Brien and Elaine G Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010.
21. Terry C Davis, Sandra W Long, Robert H Jackson, EJ Mayeaux, Ronald B George, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Family medicine*, 25(6):391–395, 1993.
22. Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, et al. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522, 2005.
23. Rafal Kocielnik et al. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pages 881–894. ACM, 2018.
24. Rafal Kocielnik and Gary Hsieh. Send me a different message: Utilizing cognitive space to create engaging message triggers. In *CSCW*, pages 2193–2207, 2017.
25. Yuta Katsumi, Suhkyung Kim, et al. When nonverbal greetings make it or break it: the role of ethnicity and gender in the effect of handshake on social appraisals. *Journal of Nonverbal Behavior*, 41(4):345–365, 2017.