Design and Implementation of Conversational User Interfaces for Health

Raina Hope Langevin

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Gary Hsieh, Chair
Andrea Hartzler
Julie Kientz
Aaron Lyon

Program Authorized to Offer Degree:
Human Centered Design and Engineering

University of Washington

Abstract

Design and Implementation of Conversational User Interfaces for Health

Raina Hope Langevin

Chair of the Supervisory Committee:

Associate Professor Gary Hsieh
Human Centered Design and Engineering

Conversational user interfaces (CUIs) have the potential to support users across varied health domain areas. Yet barriers remain to the implementation and adoption of CUIs, such as lack of trustworthiness and consideration for cultural context. There exist a number of conversational-based interventions that have been proven to be usable and effective in addressing health conditions. However, there are still few conversational-based health interventions implemented and studied in real-world settings. My dissertation research examines the design and implementation of CUIs for health interventions, where I integrate human-centered design (HCD) and implementation science methods to understand stakeholder needs and optimize implementation strategies. Drawing from methods and frameworks in HCD and implementation science, I explore CUI design considerations and challenges in two health contexts: 1) social needs screening in a large public hospital emergency department, and 2) breast cancer screening outreach for Black/African American women.

First, I adapt human-centered design approaches to improve the design of CUIs by developing and validating heuristics for conversational agents. I conduct this research to demonstrate that conversational user interfaces require unique design guidelines and considerations. Second, I deploy a conversational user interface in a real-world context to evaluate its effectiveness in supporting patient engagement. Through such deployment, I highlight implementation challenges to integrating a social needs screening chatbot into an emergency care setting, as well as individual, contextual, and intervention-related factors that may in-

fluence engagement in CUI interventions. Lastly, I engage in multidisciplinary design work to integrate methods from HCD and implementation science to improve CUI design for health interventions. I describe the development of a chatbot intervention aimed to facilitate breast cancer screening outreach. In summary, my dissertation demonstrates that adaptation of established usability heuristics for conversational user interfaces can lead to improved usability and engagement. I also discuss how an integrated approach of human-centered design and implementation science methods may combine the strengths of both disciplines in the design of chatbot implementation strategies. My dissertation makes (1) methodological contributions through the development of usability heuristics, (2) artifact contributions through the development of CUIs and through the design recommendations that arise through real-world deployments, and (3) empirical contributions through studying how CUI design components relate to engagement.

Acknowledgements

First and foremost, I would like to thank my advisor, Gary Hsieh, for his unwavering guidance and support throughout my graduate research. I am grateful to Gary for being my advisor, for always being willing to give feedback, share advice and engage in insightful research discussions. Gary has provided unparalleled mentorship in conducting research and navigating interdisciplinary collaboration.

I am fortunate to have four distinguished members on my committee, Andrea Hartzler, Julie Kientz, Aaron Lyon and Carmen Gonzalez, who I am thankful for their inspiring and thoughtful feedback and critique on my dissertation research. I am grateful as well to many research and teaching mentors: Beth Kolko, Chirag Shah, Irini Spyridakis, Sean Munson, Mark Zachry, Jennifer Turns, Kristin Dew, and Jason Yip.

This dissertation would not have been possible without many great colleagues and collaborators whom I had the chance to work with. I want to thank Herbie Duber, Leah Marcotte, and Bridgette Hempstead for welcoming me into their collaborations and partnerships to address health disparities. I am fortunate to have gotten to know them and share in their passion to advocate for people's health. I want to thank Callan Fockele, Dennis Hsieh, Victoria Fang, Anisha Ganguly, Rachel Yung, Chantal Cayo, Paula Houston, Nkem Akinsoto, Nidhi Agrawal, Nora Henrikson, and Predrag Klasnja for their valuable support and feedback on our research projects at UW Medicine. The emergency department staff and study coordinators Thomas Paulsen, Kyle Steinbock, and Layla Anderson were instrumental in facilitating the study recruitment. I would also like to acknowledge the Digital SDoH Workgroup for their feedback throughout the HarborBot project, Rafal Kocielnik for the alpha development of HarborBot, and Scott James George, Harpreet Singh and the team at Harbor-UCLA Medical Center for their support.

I would like to thank all of the participants for sharing their stories and taking the time to talk with

me. I want to thank the many incredible undergraduate and graduate collaborators that I worked with and who assisted with this research: Andrew Berry, Jinyang Zhang, Amelia Wang, Georgia Kenderova, Pranuti Kalidindi, Aiza Ali, Natasha Schmid, and Xuan Song. This research was even more enjoyable with their company and I am grateful for their support. I am grateful to Ross Lordon, Benjamin Cowan, Tad Hirsch, and Thi Avrahami, for the many discussions about conversational user interface design and for being a sounding board during the early years of my PhD. I greatly appreciate my mentors during my internships, Daniel Strazzulla Ortega, Christiana von Hippel, Andrés Monroy-Hernández, and Yu Jiang Tham, who supported my growth as a researcher and allowed me to explore new research topics. My first introduction to research began in the ROC HCI lab at the University of Rochester, and I am thankful to Professor Ehsan Hoque for giving me an opportunity to delve into research.

I would not be where I am today without the endless support and the love I received from my family. Words cannot express how grateful I am for being surrounded by my amazing family. To my loving husband, thank you for your support, wisdom and advice in all aspects of life. I feel lucky to have you by my side all these years as my best friend. You have made each moment a lot more beautiful.

While there are too many to name, I am grateful to my friends within and outside the research community for their encouragement. I am grateful to my PhD cohort for always being sources of positivity and encouragement during this PhD journey: Kenya Mejia, Michael Beach, Calvin Liang, Hannah Twigg-Smith, Rafael Silva, Susanne Kirchner-Adelhardt, Regina Cheng, Yihan Yu, Steven Goodman, Melinda McClure Haughey, Brian Kinnee, Burren Peil, and Joshua Vasquez. I would also like to acknowledge the members of Lab 325 and the Prosocial Computing Lab: Rafal Kocielnik, Mia Suh, Elena Agape, Lucas Colusso, Jenna Frens, Spencer Williams, Himanshu Zade, Keri Mallari, Lubna Razaq, Ruoxi Shang, Ruican Zhong, Donghoon Shin, Akeiylah DeWitt, Jay Cunningham, Emma McDonnell, and Sam Kolovson. Thank you all for the conversations and feedback during many research discussions, study pilots, and practice talks.

This research could not have been possible without funding and support from the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319, and the National Cancer Institute of the National Institutes of Health under Award Number P50CA244432 and by grant number K12HS026369 from the Agency for Healthcare Research and Quality.

Dedication

to my family

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Conversational user interfaces (CUIs) are dialogue systems that recognize the users' text or speech, manage the interaction, and convey information back to the user [Glass, 1999]. In the last decade, CUIs have continued to grow in popularity, through the uptake of text and voice-based conversational systems such as chatbots and Intelligent Personal Assistants (IPAs) respectively. Recently, with the advent of large language models (LLMs), conversational interactions are becoming even more integrated into our daily lives and workflows. CUIs have the potential to support users across varied health domain areas and improve health behavior [Laranjo et al., 2018; Montenegro et al., 2019]. There is evidence that text messaging has positive effects on behavior change across demographic differences [Cole-Lewis and Kershaw, 2010]. Further, conversational agents may be effective messengers for facilitating patient engagement at an individual level [Kocielnik et al., 2021]. CUIs may support more efficient workflows in healthcare contexts as an alternative to patient outreach, screening and referral. However, many challenges remain to both designing and implementing CUIs in practice. There are numerous factors to consider during the design process, such as usability, efficiency, engagement, trust, and design decisions regarding the CUI's voice, style of language, and personality [Kocielnik, 2021]. Additionally, conventional methods for assessing usability and user experience are not sufficient and accurate when applied to chatbots [Holmes et al., 2019]. Further, even though conversational users interfaces have been proven to be an effective mode of communication, they face challenges with uptake and long term use. In intervention development, there is work required to reflect on a number of implementation considerations [Greenhalgh et al., 2017] and to build shared vision among stake-

holders, engage staff, enact new practices, and monitor impact. Implementation and sustainment are rarely considered during CUI design. The majority of studies on CUIs are done in controlled environments rather than the context of use [Kocaballi et al., 2022]. As a result, we do not have a clear picture of implementation considerations, such as how CUIs might fit into existing workflows. It has been recognized as well that there is a lack of consensus on measures and definitions [Rapp et al., 2021], and that we need more research on the effects of chatbot design features (e.g., visual appearance, personality) in long-term and daily life settings [Ter Stal et al., 2020]. We still have a limited understanding about which design aspects of conversational user interfaces lead to effective outcomes in different contexts [Abd-Alrazaq et al., 2019]. Research on the implementation of conversational technologies in healthcare would benefit from research approaches that are informed by stakeholder values, culture, and preferences.

In recent years, the integration of additional fields and research approaches has been proposed as a way to address the "research-practice gap" and more rapidly translate research into action [Unertl et al., 2016]. One promising approach to improving the fit between health interventions and contexts is to combine methods from human-centered design (HCD) and implementation science to better identify determinants of intervention success, develop implementation strategies and evaluate outcomes [Dopp et al., 2020; Lyon and Bruns, 2019; Lyon et al., 2020]. Human-centered design is an approach to problem-solving that places people (direct and indirect stakeholders) at the center of the development process. It is often, but not always, applied to the design of technology to improve usability by focusing on users' needs. Implementation science is the scientific study of methods to promote the systematic uptake of research findings into routine practice. Implementation science may help human-computer interaction (HCI) researchers and designers to understand how to best use specific interventions and strategies that have been proven to work in similar settings. There has been recent attention by HCI researchers to call for systemic changes in design research practice [Howell et al., 2021]. As noted by Howell et al., "when we focus too tightly on the feasibility of a particular device, we miss out on accounting for how our design work matters (or does not) within a particular context." These efforts to understand why design projects may fail, through more open reporting of contextual factors and ongoing maintenance, may be particularly aligned with implementation science with its focus on context and sustainment of interventions. Implementation science could provide guidance at the start of the research planning, and in later stages by rigorously evaluating potential interface components.

Within the implementation science field as well, there have been efforts to rethink research practices. Implementation scientists have proposed integrated frameworks that apply human-centered design, to enhance usability and improve contextual appropriateness. These integrated approaches include both linear (e.g., IDEAs) and non-linear frameworks (e.g., Discover, Design/Build, and Test (DDBT) framework, Accelerated Creation-to-Sustainment (ACTS) model) [Mummah et al., 2016; Lyon et al., 2019; Mohr et al., 2017]. In early design stages, HCD may provide guidance on improving usability as a way to engage prospective users to understand their needs and determinants to health outcomes. By attending to usability and user experience, this can make traditional intervention protocols easier to learn and more accessible for stakeholders [Lyon and Bruns, 2019]. Another important consideration is to understand how interventions are actually developed in practice. We don't know the effectiveness of different research approaches and frameworks, which are recently starting to be applied in research projects [Hartson et al., 2022]. Implementation strategies — "methods or techniques used to enhance the adoption, implementation, and sustainability of a clinical program or practice" [Proctor et al., 2013] — are central to implementation science. As implementation science has evolved, experts have recognized that the number of implementation strategies has grown [Powell et al., 2015], and developing or tailoring implementation strategies may benefit from integrating approaches from other disciplines (e.g., behavioral economics and human centered design) [Haines et al., 2021; Beidas et al., 2021; Lyon et al., 2019]. Yet, current guidance on how to effectively incorporate methods from other disciplines to develop and refine innovative implementation strategies is limited. While most implementation strategies target service providers of the contexts in which they work, CUIs can function as a consumer-facing implementation strategy. CUIs may be introduced as interactive strategies which are delivered by providers and target individuals' motivation and capability to engage with the CUI intervention, similar to capacity-building and dissemination strategies targeted towards service providers [Leeman et al., 2017]. CUIs may function as consumer-facing implementation strategies by directing certain strategies, such as "developing educational materials" and "reminding," towards consumers [Powell et al., 2015]. In my dissertation, I aim to extend this literature through lessons learned in practice from integrating human-centered design and implementation science towards the improved design and development of CUIs.

1.1 Thesis Statement

In my research, I seek (1) to build upon general usability guidelines to improve the design of conversational user interfaces by developing and validating heuristics for conversational agents, (2) to deploy and evaluate chatbot interventions to understand design considerations and challenges in real-world contexts, and (3) to integrate human-centered design and implementation science methods in a rigorous approach to the design of chatbot interventions to improve health outcomes.

My dissertation research, therefore, demonstrates the following thesis statement:

> *Engagement with conversational user interfaces for health can be improved by integrating human-centered and implementation science methods to understand stakeholder needs and optimize implementation strategies.*

I propose that using human-centered can reveal stakeholder needs, while implementation science methods may help to prioritize intervention components for the specific health activity and context. I also reflect on how a contextual understanding can support the appropriate design and adaptation of conversational user interface interventions aimed at improving engagement in health contexts.

1.2 Dissertation Overview

My dissertation has spanned three key research activities across two health contexts in support of my thesis. The demonstration of my thesis follows research phases that are typically seen in the human-centered design process [Maguire, 2001; Harte et al., 2017; Melles et al., 2021], and the implementation science process to guide the design of interventions [Collins et al., 2005; Broder-Fingert et al., 2019; Packel et al., 2019]. I propose that these research phases are complementary and can be integrated to support the design and implementation of health interventions.

1. **Exploration:** Within the exploration phase, the goal is to draw from human-centered methods in the "Defining and understanding" and "Designing" phases to engage stakeholders, understand their needs, and to begin designing potential implementation strategies. Here, we use design prototypes and qualitative methods to understand participants' facilitators and barriers to healthcare screening.

We synthesized knowledge in a rapid evidence review and integrated human-centered methods into the Preparation phase of the Multiphase Optimization Strategy (MOST) to develop a conceptual model for a chatbot intervention. We leverage the persuasive health message framework [Hall and Johnson-Turbes, 2015; Witte et al., 1995] and trust in human-robot interactions [Siau and Wang, 2018] to gain an understanding of how a chatbot intervention might facilitate trust and engagement.

2. **Designing & Adapting:** Within the designing and adapting phase, I draw from human-centered methods in the "Adapting" phase to propose adapting usability heuristics for conversational user interfaces. I also iterate on the chatbot design by testing and adapting components of chatbot interventions for improved trust and engagement. Here, I draw from the MOST framework [Collins et al., 2005; Broder-Fingert et al., 2019] Optimization phase to identify the most effective chatbot design. I identify chatbot persona and messaging as design components that can be modified to improve stakeholder engagement.

3. **Evaluation:** In the evaluation phase, the goal is to draw from human-centered methods in the "Deploying and evaluating" phase to conduct a deployment and interview study, and implementation science measures and frameworks in the "Evaluation" phase to identify opportunities for continuous adaptation of interventions. Here, I deploy chatbot systems for user studies and analyze the data from the deployments to propose design considerations. We employ the use of implementation outcome measures, and the Consolidated Framework for Implementation Research (CFIR) [Damschroder et al., 2009] in the analysis of interview data. We used concurrent triangulation as a mixed methods approach to confirm findings within our study [Creswell and Creswell, 2017].

My dissertation draws from human-centered design and implementation science methods and frameworks to understand the needs of stakeholders within health contexts, and to optimize the design of conversational user interfaces as an implementation strategy. In Chapter 3, I adapt human-centered design guidelines for the evaluation of conversational user interfaces to create new usability heuristics that address specific design considerations related to conversational systems. I apply findings from the development of these usability heuristics to improve the design of a chatbot for social needs screening and resource provision. I then deployed the chatbot and conducted mixed-methods research to evaluate the chatbot implementation

in a hospital emergency department (ED) (Chapter 4). In Chapter 4, we study the implementation of a social needs screening chatbot and report validated outcome measures to demonstrate implementation success. The use of valid measures is important for monitoring and evaluating the success of implementation efforts and comparing the effectiveness of alternative implementation strategies. The goal of this research was to evaluate the chatbot situated in the context of the ED and patient workflow. Drawing from the CFIR framework, we investigated patient perceptions of the chatbot as an acceptable and appropriate modality for screening in the ED. While finding that a chatbot can address patients' trust concerns regarding other ED visitors and healthcare providers, patients may need guidance from community health workers and healthcare professionals at other steps in the intervention process. For example, patients rarely communicated their social needs responses to providers during their visit and some expressed a desire to receive follow-up from an individual after completing the screening. Chapter 4 raised further questions on how to design CUIs for sustained engagement. In particular, it is unknown how users will engage with chatbot screening when there are no immediate benefits or they face competing demands outside of the ED waiting room.

In Chapter 5, we use human-centered and implementation science methods to design a chatbot for breast cancer screening outreach for Black/African American women. In this work, we engaged community partners and stakeholders in the design process through interviews and focus groups to gain insights about the potential adoption of a chatbot with culturally relevant health education. We draw from the Multiphase Optimization Strategy (MOST) framework, a framework for developing efficacious, efficient, scalable and cost-effective interventions, to assess the performance of chatbot intervention components and their interactions [Collins et al., 2005; Packel et al., 2019]. Multiple studies have utilized the MOST framework to develop and test intervention components [Broder-Fingert et al., 2019]. MOST is a method that involves three phases (Preparation, Optimization, and Evaluation) for systematically building and evaluating interventions to ensure they comprise active components [Collins et al., 2005; Broder-Fingert et al., 2019]. We used the MOST framework and engaged in an Exploratory phase to (1) center stakeholders and community partners in a qualitative analysis to understand breast cancer screening determinants, and (2) to develop a conceptual model for the intervention, identify core components, and determine what outcomes should be optimized, drawing from the MOST Preparation phase. We then engaged in an Optimization phase to conduct a randomized factorial experiment of specific components, and used convergent mixed-methods to analyze the

data and suggest an optimized delivery strategy. We demonstrate how methods from human-centered design and implementation science may complement each other, as most implementation evaluation processes include mixed qualitative and quantitative measures [Bauer et al., 2015]. Human-centered design approaches may be helpful in the early stages to guide design with the field's focus on novelty and iterative design. Following this, implementation science may support human-centered design in later stages with its focus on delivery as intended and continuous adaptation.

My dissertation contributes an empirical understanding of the current challenges of conversational user interface design, and proposes the adaptation of human-centered design guidelines. My dissertation makes (1) methodological contributions through the development of usability heuristics, (2) artifact contributions through the development of CUIs and through the design recommendations that arise through real-world deployments, and (3) empirical contributions through studying how CUI design components relate to engagement. I design and evaluate chatbots in two health contexts by integrating human-centered design and implementation science methods: 1) social needs screening in a large public hospital ED and 2) breast cancer screening outreach to support Black/African American women in scheduling mammograms. In the following sections, I will discuss related work on the design of conversational user interfaces. I will present my research on the adaption of usability heuristics for conversational user interfaces. Building on my prior work, I will describe existing and ongoing work to implement conversational user interfaces in real-world contexts using human-centered design and implementation science approaches.

Chapter 2

Background

My dissertation research explores the challenges and barriers to designing conversational user interfaces for health interventions aimed at improving engagement. In this chapter, I provide an overview of conversational user interface design and the need for adaptations (Section 2.1). I then describe the challenges of conversational user interface design for health contexts (Section 2.2), and approaches towards integrating human-centered design and implementation science methods to improve the design of technology for health interventions (Section 2.3).

2.1 Overview of Conversational User Interfaces

Conversational user interfaces (CUIs) are dialogue systems with a wide range of applications. At minimum, a dialogue system is intended to recognize the users' text or speech, manage the interaction, and convey information back to the user [Glass, 1999]. Depending on the domain, a conversational agent may be designed for entertainment, companionship, informational or task-based purposes. Conversational user interfaces, which may be referred to as conversational agents, can have different modalities, including text, speech and multimodal embodiment. Examples of conversational agents include well-known text-based conversational agents, such as ALICE, and speech-based conversational agents, such as Alexa, Siri and the Google Assistant. While there is broad guidance on the design of CUIs, recently there have been strides towards consolidating and validating guidance in related areas, such as human-AI interaction [Amershi et al., 2019], and human-like chatbot experiences [Svenningsson and Faraon, 2019].

2.1.1 Value of Conversational User Interfaces for Health

Conversational user interfaces have the potential to support users across varied health domain areas and improve health behavior [Laranjo et al., 2018; Montenegro et al., 2019]. Mobile technology interventions, such as CUIs, are promising for health and screening outreach to address health disparities and address limitations of prior interventions in terms of costliness and relevance [De Jesus et al., 2021; Ruco et al., 2021]. These interventions using short message service (SMS) text are accessible to individuals across a range of sociodemographic factors [Center, 2021] and have been shown to be effective in primary care behavioral and disease management interventions [Free et al., 2013]. Conversational user interfaces could be delivered via SMS text and could provide an efficient, individualized approach to health outreach. There is evidence that text messaging has positive effects on behavior change across demographic differences [Cole-Lewis and Kershaw, 2010]. Further, CUIs may be effective messengers for facilitating patient engagement [Kocielnik et al., 2021]. These conversational systems can act as virtual health navigators providing information about and connecting individuals to healthcare [Gardiner et al., 2013]. Chatbots are computer programs that simulate conversations with users that are increasingly adopted in the healthcare field [Laranjo et al., 2018; Montenegro et al., 2019], and have the potential to address the above screening barriers. Chatbots may increase patient uptake as a more understandable and engaging tool compared to traditional online surveys [Xiao et al., 2020; Kocielnik et al., 2019, 2021]. Individuals may be more willing to disclose personal information [Lucas et al., 2014] and social needs information [Gottlieb et al., 2014] to a computer. They may feel more inclined to use conversational agents for discussing sensitive health topics, such as addiction [Auriacombe et al., 2018], depression [Philip et al., 2017], and post-traumatic stress disorder [Lucas et al., 2017], as technology can enable more confidential methods of information and support seeking [Stowell et al., 2018; Cornelius et al., 2012]. Chatbots have been shown to increase levels of trust in web-based information [Rickenberg and Reeves, 2000], and may also be easy to use and scale [Bickmore et al., 2010].

Despite growing interest in conversational-based technology for health, there are few published studies on conversational user interfaces in healthcare [Laranjo et al., 2018; Kocaballi et al., 2022; Tudor Car et al., 2020; Milne-Ives et al., 2020]. There is little data on integrating conversational agents in primary care outreach and none that we are aware of that specifically address healthcare disparities [Graham et al., 2020]. The literature on conversational agents in healthcare is largely aimed at treatment and monitoring of health

conditions, such as mental health [Kamita et al., 2019; Ly et al., 2017; Fitzpatrick et al., 2017; Inkster et al., 2018], Alzheimer's [Griol and Callejas, 2016], heart failure [Galescu et al., 2009], asthma [Rhee et al., 2014], and HIV [Vita et al., 2018], health service support, such as patient history taking [Denecke et al., 2018; Ni et al., 2017] and triage and diagnosis support [Razzaki et al., 2018; Ghosh et al., 2018], and patient education, on topics such as sexual health [Wilson et al., 2017], smoking [Wang et al., 2018], alcohol use [Elmasri and Maeder, 2016], and breast cancer [Chaix et al., 2019]. Literature on digital health interventions emphasizes need for careful attention to and planning for implementation to optimize integration in the healthcare system and patient use [Parker and Harris Lemak, 2011]. Thus, we identified chatbots as a promising implementation strategy to address health disparities in social needs screening and breast cancer screening; however one that warrants contextual understanding and rigorous methods to design and tailor.

2.2 Challenges in Conversational User Interface Design for Health

Challenges remain to the design and implementation of conversational user interfaces, such as low efficiency, perceptions of artificiality and high expectations of intelligence [Kocielnik et al., 2021]. While there is an increased interest in using these technologies, designing conversational user interfaces is not easy. There are a number of barriers to interacting with conversational user interfaces, such as unmatched expectations of the system's capabilities [Clark et al., 2019], differences in conversation styles [Thomas et al., 2020], increased cognitive load for particular user groups [Wu et al., 2020] and social embarrassment [Cowan et al., 2017]. Past work has diverged on whether chatbots should exhibit human-like characteristics and a number of desirable human-like behaviors have been proposed [Svenningsson and Faraon, 2019]. For example, while small talk has been shown to be beneficial for establishing trust [Bickmore and Cassell, 2005], it may not be desired based on the context of the chatbot [Svenningsson and Faraon, 2019] or users' personal preferences [Liao et al., 2016]. Additionally, the design of voice interfaces is challenging. Users may be faced with a higher cognitive load as they should listen to and remember verbal information. Designers and developers must consider numerous factors during the design process.

Designing chatbot interventions is particularly challenging as users face limited technological and institutional trust. While chatbots have been shown to increase levels of trust in web-based information [Rickenberg and Reeves, 2000], adoption of chatbots may depend on user and chatbot characteristics. Though

individuals are receptive towards using health chatbots to find information, they may be hesitant to due issues regarding chatbots' accuracy, confidentiality and lack of ability to empathize with patients about emotional issues [Nadarzynski et al., 2019]. Further, the chatbot's persona may impact the credibility of its information. Users may be aware of message source and which impacts their behavior [Snyder and LaCroix, 2001]. [Cheng et al., 2008] found that participants gave socially desirable responses when using a handheld device to answer sexual behavior questions. Source characteristics such as credibility and trustworthiness can have different levels of influence depending on who is delivering the health information. CUI-based health interventions may need to position their efforts in collaboration with people and institutions that users already trust [Veinot et al., 2013]. For example, participants believed that medical providers and the government did not care about their well-being, but felt that community-based organizations were more up front and helpful and nurses were impartial.

While studies of conversational user interfaces in healthcare have shown moderate evidence of usability and effectiveness [Milne-Ives et al., 2020], there is a need for further exploration on the design and implementation of conversational user interfaces. User feedback on conversational agents in healthcare remains mixed, with some users expressing desire for interactivity and agent empathy, whereas others report a dislike of these qualities [Kocielnik et al., 2019, 2021; Milne-Ives et al., 2020]. More work is needed as well to evaluate and implement chatbots in real-world settings [Kocaballi et al., 2022; Fan et al., 2021]. Prior research on conversational agents in real-world settings has identified the need for providing actionable and accurate information [Fan et al., 2021], and emphasized the importance of designing for and with vulnerable populations, such as people with low health literacy, to improve chatbot understandability [Kocielnik et al., 2019]. The evidence base for conversational user interfaces related to mental health is sparse but growing. Recent studies have assessed the effectiveness and safety of using chatbots in mental health through randomized controlled trials (RCTs) [Abd-Alrazaq et al., 2020]. However, these studies have compared conversational user interfaces against treatment as usual [Burton et al., 2016; Fitzpatrick et al., 2017; Fulmer et al., 2018]. We still have a limited understanding about which components of conversational user interfaces lead to effective outcomes in particular contexts. The design and implementation of conversational user interfaces remains difficult, due to the large amount of design and contextual factors. Understanding how conversational user interfaces for health are used in real-world contexts and why they are effective, or

not, is essential to promote translation of digital health interventions into practice.

2.3 Harmonizing Human-Centered Design and Implementation Science

Implementation science is the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practice into routine practice [Eccles and Mittman, 2006]. The field of implementation science strives to address the "research-practice" gap between what should be done based on existing evidence and what is done in practice. Guidance has been proposed to enhance the development of health interventions through numerous frameworks [Hawkins et al., 2017; Greenhalgh et al., 2017; O'Cathain et al., 2019]. These works have encouraged combining methodologies across fields, such as design, behavioral science, and public health, to improve knowledge regarding causal mechanisms of interventions. Recent work has developed frameworks to integrate human-centered design and implementation science methods to improve usability as a determinant to the implementation, and the overall implementation of health interventions [Mohr et al., 2017; Lyon et al., 2019]. Haines et al. also proposed leveraging user-centered design (UCD) to better tailor implementation strategies to evidence-based practices (EBPs) and contexts. The use of UCD may be well suited to harmonizing EBPs, contexts and implementation strategies because its primary goals (usability and usefulness) are key determinants of perceptual (acceptability, feasibility, appropriateness) and behavioral implementation outcomes. The core tasks of implementation science include 1) identifying barriers and facilitators (i.e. determinants) of intervention success, 2) developing strategies to address determinants, and 3) evaluating implementation outcomes [Lyon and Bruns, 2019; Lyon et al., 2020]. Implementation tasks can be aided by specific approaches from HCD, and vice versa human-centered design may be informed by implementation science. Human-centered design provides methods to involve end users that can help accomplish co-creation goals, such as a focus on local knowledge, ongoing collaboration and building trust, which are theorized to enhance implementation outcomes [Proctor et al., 2011]. This may address a limitation of implementation science methods which might inadvertently attempt to design innovations for all potential users, and may not meet the needs of a particular group [Lyon et al., 2020].

Implementation science methods may also complement human-centered design in rigorously evaluating potential interface components to navigate "feature creep" [Haines et al., 2021]. Klasnja et al. described how

"usable evidence" can address the limitations of current HCI evidence base and proposed a research process that can be used to generate such evidence, through causal pathway mapping and testing of individual intervention components [Klasnja et al., 2017]. Intervention component refers to a piece of functionality of a technology that enacts a specific intervention idea for supporting behavior change. The causal pathway diagram (CPD) is an implementation science method that can be used to support development and refinement of implementation strategies and intervention components. CPDs help researchers to clarify causal assumptions and understand implementation strategies as they are intended to work [Bleijenberg et al., 2018; Lewis et al., 2018]. In building a causal pathway diagram, researchers identify the implementation strategy, the mechanism(s) through which the strategy is thought to lead to the intended outcome, proximal outcomes which may provide signals of effect earlier than the intended outcome, and the distal or intended outcome. CPDs also include moderators which may impact pathway effect and pre-conditions which are necessary for the pathway to proceed. In developing implementation strategies, causal pathway diagrams may help investigators map out implementation strategies, determinants (e.g., barriers or facilitators) addressed by the strategies, and mechanisms by which strategies are hypothesized to effect change. Human-centered design methods may benefit from the use of measures, frameworks and methods from implementation science, such as CPDs and the Multiphase Optimization Strategy (MOST) framework, in the design and refinement of technology-based implementation strategies. Recently, MOST has been extended in frameworks that utilize design processes and implementation science, such as the design and evaluation of digital health interventions (DEDHI) framework and DDBT/MOST approach, which integrates the Discover, Design/Build, and Test (DDBT) framework into the Preparation phase of MOST [Kowatsch et al., 2019; O'Hara et al., 2022]. An initial phase of human-centered design is to establish the context of use and understand users' needs [Harte et al., 2017]. In this exploratory phase, researchers often collect and analyze qualitative data through interviews, focus groups, or co-design sessions. The practice of establishing context of use and user requirements in HCD may assist with specifying implementation strategies and developing models in the implementation process [Lyon et al., 2020]. HCD methods can be used to build and inform CPDs through gaining understanding of barriers and facilitators to the implementation strategy, and key determinants to the intended outcome within a specific context. Theory and existing evidence are typically used to construct CPDs [Lewis et al., 2022], and early qualitative data from human-centered design methods and identifica-

tion of determinants can help to inform the use of theory in modeling causal pathways. The integration of human-centered design and implementation science approaches can further help to test assumptions and prioritize components of an implementation strategy.

Chapter 3

Adapting human-centered design for conversational agent evaluation

In this chapter, I discuss the adaptation of human-centered design guidelines for the evaluation of conversational user interfaces (CUIs). I propose and validate a set of 11 heuristics for conversational agents that can be generalized to text, voice and multi-modal conversational agents [Langevin et al., 2021]. Recently there have been strides towards consolidating and validating guidance in related areas, such as human-AI interaction [Amershi et al., 2019], and human-like chatbot experiences [Svenningsson and Faraon, 2019]. Unlike other forms of human-computer interfaces, there is little consensus as to best practice for the design of conversational agents [Clark et al., 2018]. Our work looks to build upon recent efforts [Murad et al., 2019][Wei and Landay, 2018], to develop a comprehensive set of heuristics for conversational agent based interactions. One common strategy to facilitate the design of technologies has been the use of formative evaluation techniques and cognitive walkthrough. These techniques can be used by designers and developers in early stages of design to eliminate usability problems. One such example is heuristic evaluation [Nielsen and Molich, 1990], a widely used discount usability testing method that identifies usability issues within a human-computer interface. In heuristic evaluation, a small set of evaluators independently examine an interface and compare its dialogue elements to a list of recognized usability principles ("heuristics"). It is an informal method that can be performed by non-experts. As a low-cost, efficient method of conducting usability evaluations, heuristic evaluation is a valuable tool for designers. Our research takes the

approach of using Nielsen's heuristics [Nielsen and Molich, 1990] as a foundation upon which to build, adapting these for conversational agent based interaction. With the additional types of interactions afforded by conversational agents, one empirical question arises: How well do the existing heuristics apply to the design of conversational agents? Can we develop a set of heuristics that are more applicable and useful for conversational agent interface design? In this chapter, we focus on validating and adapting Jakob Nielsen's 10 usability heuristics to conversational agents. In this study, I demonstrate that human-centered design guidelines - such as usability heuristics - can be adapted to improve the design of new technologies. Using conversational agent heuristics, evaluators can be better equipped to identify usability problems related to CUIs, such as context preservation, dialogue elements that create a smooth conversation, and help and guidance during the conversation. While the conversational agent heuristics do not address specific design questions (e.g., how should the chatbot greeting start?) or questions of how to implement CUIs in practice, they provide guidelines on what design standards may lead to a more usable interface. This is important as designers and developers face an overwhelming amount of design considerations and tradeoffs. The adaptation of heuristics for conversational agents provided insights not only into how to design more usable CUIs, but how to design more effective communication and create CUIs that are truthful, informative, relevant, and clear, through following cooperative principles of communication.

We sought to expand on Nielsen's heuristics using a four phased design process.[1] We first developed a set of heuristics for the design of conversational agent interfaces using prior research findings as well as our own experiences in developing these interfaces. Second, we presented these heuristics to nine experts in conversational agent design and heuristic evaluation, and incorporated their feedback. In the third phase, we evaluated our heuristics on two interfaces, a voice assistant on the Amazon Echo and an online chatbot. We compared our heuristics with Nielsen's heuristics to observe their effectiveness in identifying usability issues with conversational agents. After finding that the conversational agent heuristics performed well on the voice interface, but not the chatbot interface, we further iterated on the heuristics. Finally, in the fourth phase, we validated our heuristics on the chatbot interface by comparing them to Nielsen's heuristics. From this, we determined that the conversational agent heuristics performed more effectively than Nielsen's heuristics. We found that four evaluators can identify more usability issues when using heuristics for con-

[1] In this work, references to "we" refers to our work with collaborators including researchers and practitioners working on the design and development of conversational user interfaces.

versational agents compared to Nielsen's usability heuristics. These results were consistent with past work indicating that adapting Nielsen's heuristics is an effective method. We found that conversational agent heuristics are useful for highlighting issues related to dialogue content, interaction design, help and guidance, and trustworthiness. For example, two of the top evaluators indicated that Nielsen's usability heuristics were not applicable to the chatbot interface and created their own heuristics (such as "the wording of the chatbot dialogue" and "lack of confidentiality"). The findings from this study contribute to an empirical understanding of the current challenges and approaches to conversational agent design, and propose validated design guidelines for improved usability, which may ultimately improve the overall implementation of the CUI as an intervention [Mohr et al., 2017; Lyon et al., 2019].

3.1 Adapting Nielsen's Heuristics

Heuristic evaluation commonly relies on the set of 10 heuristics established by Jakob Nielsen [Nielsen and Molich, 1990]. Heuristics are a well-established set of guidelines that tend to result in good interface design when they are incorporated into the design process. In the 1990s, Nielsen and Molich classified usability problems of a telephone index system into nine heuristics [Molich and Nielsen, 1990]. The heuristics were based on their experiences and were supported by the principles outlined in [Inc, 1987] for the Apple desktop interface. The following heuristics were updated by Nielsen in 1994 and are still widely used today:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design

9. Help users recognize, diagnose and recover from errors

10. Help and documentation

Since Nielsen and Molich developed the initial usability guidelines in 1990, user interfaces have continued to evolve. In particular, the development of conversational agents has grown substantially with the advancement of natural language processing (NLP) and deployment of voice-enabled personal assistants and chatbots. User interface design has shifted from a focus on task-oriented, graphical user interfaces (GUI) and strides have been made towards incorporating personal engagement, and voice and speech recognition. Researchers have recognized the need to adapt Nielsen's broad set of heuristics to specific interfaces. For example, there is a wide range of heuristics available for mobile and web designers [Instone, 1997] [Chisnell et al., 2006] and past work has had success in extending Nielsen's heuristics for smartphones [Calak, 2013], ambient displays [Mankoff et al., 2003], and medical devices [Zhang et al., 2003].

There have been recent developments towards heuristics for specific modalities, like voice interactions [Wei and Landay, 2018][Murad et al., 2019]. However, we are not aware of a comprehensive set of heuristics. Due to the lack of validation for design heuristics in specific domains [Hermawati and Lawson, 2016], it is important to validate proposed heuristics in line with previous work[Amershi et al., 2019]. In this chapter, we utilize a similar design process used in prior work to develop heuristics for ambient displays [Mankoff et al., 2003]. We conduct a four phased design process as referenced in Table 3.1.

Phase 1: Heuristic Generation
Phase 2: Expert Review
Phase 3: Validation through Heuristic Evaluation
Phase 4: Validation of Revised Heuristics

Table 3.1: The four phased design process.

3.2 Phase 1: Heuristic Generation

We first conduct a literature review to consolidate guidelines and establish an initial set of 13 heuristics for designing conversational agents (see Table 2).

3.2.1 Consolidating Guidelines

We conducted a literature review and gathered 56 papers related to the evaluation or design of conversational agents. First, we searched the ACM digital library and selected 34 papers relevant to the following search terms: "evaluation of" or "guidelines" + "conversational agents," or "voice assistants". We also searched the references of the selected papers and "cited by" papers on Google Scholar and compiled a set of 22 papers. The papers spanned the years between 1977 and 2019. We then developed a list of guidelines based on 131 design suggestions from the literature. We sorted each of the design suggestions under Nielsen's heuristics and created new groups for suggestions that did not relate to the heuristics. There were none that were grouped under *Consistency and standards*.

3.2.2 Co-developed Set of Heuristics

We adapted Nielsen's heuristics and created a set of 13 heuristics based on the guidelines from literature. In a series of revisions, we iterated on the developed set of heuristics. We edited the heuristics to be less focused on visual feedback associated with GUIs. Nielsen's heuristics were also expanded to include *Clarify Capabilities*, *Context Preservation* and *Privacy*.

Through our search we also found useful unpublished research that adapted Nielsen's heuristics to evaluate a patient-centered common surgery question chatbot [Lordon, 2019]. Therefore, in the last iteration of revisions, we merged our set of heuristics with the adapted set in [Lordon, 2019]. We did not include elements of the set that were specific to health information seeking context. After we merged the sets, three authors reviewed the heuristics to provide feedback.

Inspired by [Lordon, 2019] we also included Grice's Cooperative Principles [Grice, 1975] so as to strengthen the focus on conversation between the user and the conversational agent. Grice's Cooperative Principle dictates that communication is characterized by cooperative efforts between conversational participants [Grice, 1975]. The Cooperative Principle can be understood through four maxims: quality, quantity, relevance and manner. Cooperation between conversational partners is facilitated by the quality, or truth, of what we say, the quantity of information that we provide, the relevance of what we contribute, and the clear and brief manner of our communication. The Cooperative Principle has already been applied to conversation design in dialogue systems, such as for Google Assistant [Google, n.d.].

Phase 1	Rel.	Phase 2
Visibility of system status	3.7	**Visibility of system status**
Clarify capabilities	4	**Clarify capabilities**
Match between system and the real world	4.1	**Match between system and the real world**
User control and freedom	4	**User control and freedom**
Consistency and standards	4.3	**Consistency and standards**
Error prevention	3.9	**Error prevention**
Recognition rather than recall	3.8	**Learnability**
Domain specific flexibility and efficiency of use	3.8	**Multimodal flexibility and efficiency of use**
Aesthetic, minimalist and engaging design	4.1	**Aesthetic, minimalist and engaging design**
Help and documentation	2.7	
Context preservation	4	**Context preservation**
Privacy	4.1	**Trustworthiness**
Veracity	3.8	
	N/A	**Help users recognize, diagnose and recover from errors**

Table 3.2: The conversational agent heuristics developed in Phase 1, the average relevance rating for each heuristic, and the heuristics developed in Phase 2.

We aligned the four maxims with seven of our heuristics. We matched the maxim of quantity to *Recognition rather than recall* and *Aesthetic, minimalist and engaging design*, the maxim of relevance to *Context preservation*, and the maxim of manner to *Match between system and the real world*, *Consistency and standards* and *Recognition rather than recall*. We found that maxim of quality fit under *Clarify capabilities* and *Privacy*, yet neither fully encapsulated the characteristic of "being truthful." As a response, we explicitly outlined the maxim of quality by creating the heuristic *Veracity*.

We included [Lordon, 2019]'s adaption to *Visibility of system status*, which we had not adapted initially. We removed phrases that suggested specificity to task-oriented conversational agents, as well as references to "visual or audible" system responses in [Lordon, 2019]'s set that were targeted towards smartphone modalities. The only heuristic that remained without adaptation was *Help users recognize, diagnose and recover from errors*.

3.3 Phase 2: Expert Review

After generating the heuristics in Phase 1, an expert evaluation was conducted to gather feedback on the modified heuristics developed. In the expert evaluation, participants were presented with a list of heuristics and asked to rate and comment on their relevance to the evaluation of conversational agents. This study received Institutional Review Board (IRB) approval for Phases 2, 3 and 4.

3.3.1 Participants

We recruited participants by contacting individuals in our professional network and providing them with an introduction letter and a link to the study. We included participants who fit the following inclusion criteria: adults over the age of 18, and having work experience in conversational agent design and usability testing methods. Participants were informed that they were identified to participate as they have expertise in the areas of conversational agent design and usability testing methods.

Five researchers, two professors, one user interface designer, and one digital initiative leader participated in our evaluation. The average self-rated level of experience with heuristic evaluation was 3.1 and experience with conversational agent design was 4.2 on a 5 point Likert scale (5 being the highest, 1="never heard of it" and 5="expert"). All of the participants had work experience designing or building conversational agents. Participants had previously designed or built 9 conversational agents on average. Additionally, participants had conducted an average of 6 heuristic evaluations. Three of the nine experts in conversational user interface design had not conducted heuristic evaluations before, which led to a reported average of 6 evaluations conducted. When not including those experts, the average number of evaluations conducted was 9.

3.3.2 Procedure

We asked participants to review the heuristics developed in Phase 1 and assign a relevance rating on a scale of 1 to 5 (5 being the highest) to indicate how relevant each heuristic was to the evaluation of conversational agents. They were encouraged to provide comments on the heuristics and were given the option to suggest additional heuristics for conversational agents as well.

3.3.3 Results

As shown in Table 3.2, the relevance ratings for each of the heuristics were above 3.7, with the exception of *Help and Documentation* with the lowest relevance rating of 2.7. One respondent said that the conversational agent should be self-explainable, rather than having the need for documentation. Based on the experts' feedback, we removed the heuristic *Help and documentation*.

Respondents also noted that while truthfulness is an important quality for gaining user trust, *Veracity* may not be a necessary usability requirement. Thus, we removed *Veracity* and included elements of the heuristic in *Trustworthiness* to reflect their comments.

Finally, we made a number of adjustments to the other heuristics. We added clarifications to *Domain specific flexibility and efficiency of use*, such as the addition of "verbal shortcuts." We also made changes to *Recognition rather than recall* to place less emphasis on visual information, and *Match between system and the real world* to encourage smooth dialogues, rather than mirroring real conversations.

3.4 Phase 3: Validation through Heuristic Evaluation

In Phase 3, we proceeded to apply the modified heuristics to two conversational agents. We conducted two studies to evaluate the effectiveness of our modified heuristics to Nielsen's original heuristics. We recruited one set of participants for an in-person study and another set to complete the study online. In each study, we used a between-subjects design where one group was asked to evaluate the conversational agent using Nielsen's usability heuristics, and the second group was asked to evaluate the same conversational agent using the modified heuristics.

We chose systems that were both in-development so that evaluators could find a number of usability issues in the heuristic evaluation. The systems were also selected to cover both text and voice modalities. We first evaluated a voice-based conversational agent, and then a text-based conversational agent.

3.4.1 Systems Evaluated

In the in-person study, we asked participants to evaluate a voice assistant using the Amazon Echo. This was structured as an in person study so we could ensure all participants had access to the same physical device,

Amazon Echo. We searched for an Alexa skill on the Amazon website that was in the Social category and had customer ratings with less than 4 out of 5 stars. This was done to ensure that the system had a sufficient number of usability issues for the heuristic evaluation. We observed that in the reviews of low-rated skills, users described a number of issues with the system that accompanied the low rating. The Social category was chosen to vary the types of systems evaluated. We searched for an interface with more free-form input, as the chatbot provided predefined options. We selected an Alexa skill that connects to a Slack workspace and can be used to read, send and react to messages. We set up a fictional Slack workspace that was linked to the Amazon Echo. Participants were given a username to communicate with other users in a university department.

In the online study, participants evaluated an in-development text-based chatbot interface. The interface was designed to collect survey information from people in hospital emergency departments. The chatbot asks users various questions regarding their health, housing situation, and employment, to screen users for unmet social needs [Kocielnik et al., 2019].

3.4.2 Participants

In-person Heuristic Evaluation

There were 16 participants recruited via Slack and email from a large university. We assigned 8 participants to each condition for the in-person heuristic evaluation sessions using the Alexa skill. The participants included 12 graduate students, two UX researchers, one engineering intern and one undergraduate student. The backgrounds of the participants ranged from computer science and engineering, user research, human-centered design, and healthcare.

In the group that used Nielsen's heuristics, the average self-rated level of experience with heuristic evaluation was 2.9 and experience with conversational agent design was 2.3 on a 5 point Likert scale (5 being the highest, 1="never heard of it" and 5="expert"). Six of the participants had conducted heuristic evaluations 1-5 times, one had conducted 6-10 evaluations and one more than 10 evaluations. In the group that used conversational agent heuristics, the average self-rated level of experience with heuristic evaluation was 2.8 and experience with conversational agent design was 2.8 on a 5 point Likert scale. Five of the participants had done heuristic evaluation 1-5 times, and three had never conducted a heuristic evaluation

before.

Online Heuristic Evaluation

We recruited 16 participants via Slack and email from our professional network for the online heuristic evaluation sessions. There were 9 participants in the group that used Nielsen's heuristics and 7 participants in the group that used the conversational agent heuristics. The participants included 10 graduate students, two students, two engineers, one researcher, and one UX design intern. The background of the participants ranged from human-computer interaction, UX/UI design, psychology, computer science, service design, archives and libraries, user research and marketing.

In the group that used Nielsen's heuristics, the average self-rated level of experience with heuristic evaluation was 2.4 and experience with conversational agent design was 2.4 on a 5 point Likert scale. Six had conducted heuristic evaluation 1-5 times and three had never conducted a heuristic evaluation before. In the group that used conversational agent heuristics, the average self-rated level of experience with heuristic evaluation was 3.1 and experience with conversational agent design was 2.7 on a 5 point Likert scale. Five participants had conducted heuristic evaluation 1-5 times, one had never conducted a heuristic evaluation, and one had conducted more than 10 evaluations. While participants in Phase 3 were skilled in heuristic evaluation on average, there was a mix of non-expert participants, who had a lower self-rated experience with heuristic evaluations, and participants with more expertise.

3.4.3 Procedure

In both the in-person and online studies, the instructions and time provided in the in-person and online contexts were the same to minimize the effect the study context. All participants read the same instructions on a Google document and the in-person participants had minimal interactions with the experimenter during the evaluation. Participants were presented with a list of heuristics (either our modified heuristics or Nielsen's original heuristics), and a description of the conversational agent and usage scenario. Participants were asked to examine the interface several times and create a list of usability issues. For each usability issue, they were told to explain the issue, reference one or more heuristics that it was related to, and assign a severity rating on a scale of 0 to 4 (4 being highest) to indicate how severely the issue limits the users'

ability to use the conversational agent. They were also permitted to include additional heuristics that related to one of the usability issues. Participants were compensated with a $25 gift card for conducting a one hour heuristic evaluation of the conversational agent.

3.4.4 Results

The authors first conducted an informal expert review to generate a master list of all known usability issues, following methodology in past work [Mankoff et al., 2003] [Nielsen and Molich, 1990]. With expertise in HCI, conversational agent interaction, heuristic development and evaluation, the authors reviewed the two interfaces and internally generated a list of usability issues. This list was then combined with all of the issues identified by participants to create the final master list of usability issues. From this list, we removed non-issues which conveyed a misunderstanding regarding the interface or did not refer to a specific usability issue. In total, there were 42 issues in the master list for the Alexa skill and 53 issues for the chatbot.

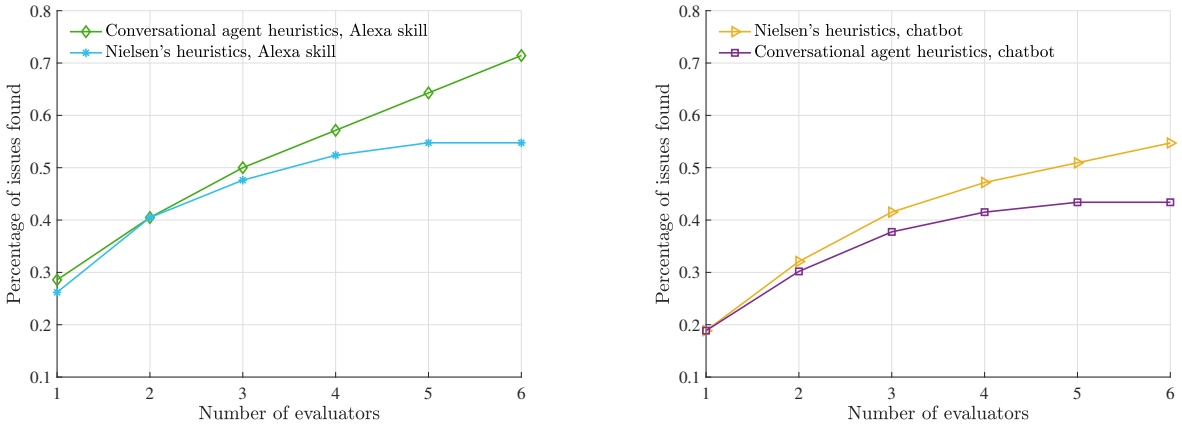
To evenly balance the number of participants and experience with heuristic evaluation in each group, we removed participants who had conducted heuristic evaluation more than six times. We then selected the top 6 evaluators in each group to compare the number of usability violations. In Table 3.3 and Figure 3.1, we refer to the 12 participants who evaluated the Alexa skill as *voice* and the 12 participants who evaluated the chatbot as *chatbot*. While four evaluators are recommended by the literature, we chose to display the top 6 evaluators in Phase 3 to show as much information as possible.

		Phase 3	
Participant set	Experts	CA	Nielsen
voice	9	30	23
chatbot	31	23	29

Table 3.3: Number of usability issues found by the experts, and the top six evaluators in the conversational agent (CA) and Nielsen groups in Phase 3.

In-person Heuristic Evaluation

While the groups were similar based on self-rated experience with heuristic evaluation, the Nielsen condition had done more heuristic evaluations in practice. We balanced the experience of the participants and selected the top 6 participants from the Nielsen group and top 6 participants from the conversational agent group



(a) *voice* **(b)** *chatbot*

Figure 3.1: Percentage of issues found by the top six evaluators using the conversational agent heuristics and Nielsen's heuristics on the two interfaces in Phase 3.

who identified the most issues from the master list of issues. We removed two participants from the Nielsen group from this selection process who had high expertise; one had conducted 6-10 heuristic evaluations and one had done more than 10 heuristic evaluations.

The results showed that the conversational agent heuristics were better able to identify issues than Nielsen's for the Alexa skill. As shown in Table 3.3, the top 6 evaluators using the conversational agent heuristics identified 30 out of 42 issues compared to those using Nielsen's heuristics 23 out of 42. In Figure 3.1a, we sort the participants by additional unique ideas found and find that the top four evaluators in the group using conversational agent heuristics found 57% of known issues, compared to the group using Nielsen's heuristics found 52% of known issues. The use of four evaluators is recommended as an optimal number needed to uncover the majority of issues [Nielsen, 1994]. As the number of evaluators increases, the conversational agent heuristics continue to uncover unique issues; six evaluators ultimately find 71% of issues using our heuristics compared to 55% when using Nielsen's.

Online Heuristic Evaluation

To balance the number of participants, we selected the top 6 participants from the Nielsen group and the top 6 participants from the conversational agent group who identified the most issues from the master list of issues. We also balanced the actual experience with heuristic evaluation and removed one participant from

the conversational agent group who was an expert in heuristic evaluation and conducted heuristic evaluation more than 10 times.

In the online heuristic evaluation, the conversational agent heuristics were not more effective than Nielsen's heuristics for the chatbot interface in the online study. In Table 3.3, the top 6 participants using our heuristics identified 23 out of 53 issues, while those using Nielsen's heuristics found 29 out of 53 issues. Nielsen's heuristics offered more coverage of usability issues for the chatbot as shown in Figure 3.1b. We found that the top four evaluators found only 42% of usability issues when using the conversational agent heuristics, compared to 47% of usability issues using Nielsen's heuristics.

While the conversational agent heuristics were more effective than Nielsen's in identifying issues with the Alexa skill, they were less effective in regards to the chatbot interface. To address the limitations of the heuristics, we revised the conversational agent heuristics for further testing with the chatbot.

3.4.5 Revisions

Based on the results of the heuristic evaluation, we made a number of revisions to the conversational agent heuristic set. We first went through violations found by Nielsen's set or by the experts, but not by the conversational agent heuristics. We then updated the conversational agent heuristics to better address these violations.

In the chatbot evaluation, we noticed that Nielsen's heuristics captured more visual design violations, such as "text is overflowing from multiple choice options". Thus, in the conversational agent heuristics, we reframed the introductory text and made explicit the terms (visual design, dialogue etc) in the heuristics to prepare them to evaluate different modalities. We removed terms such as "voice interfaces" and changed them to "interfaces" in *Aesthetic, minimalist and engaging design* to better generalize the heuristics to interfaces with multiple modalities. We re-incorporated "Follow platform conventions" to *Consistency and Standards* because one participant using Nielsen's heuristics noted an inconsistency in the colors on the checklist across mobile and web platforms.

The experts brought up two issues regarding the chatbot's audio output that were not found by the conversational agent heuristics. The experts found that the "use of voice as output is not appropriate for asking sensitive questions" and the "use of voice as output, but not input, doesn't match user expectations".

In response, we added “depending on the use context” and “input and output” to *Flexibility and efficiency of use*. Additionally, one participant in the Nielsen condition brought up an issue that the chatbot’s robotic voice was off-putting. We thus added the use of “an appropriate voice” to *Match between system and the real world*. The sentence “Make information appear in a natural and logical order” was included in Nielsen’s original heuristics, but was removed when we first iterated on the heuristics as we emphasized mirroring natural conversation at the time. We added it to our revised heuristics as "the ordering of the questions is not organized well" was a violation identified only in the Nielsen condition.

In the evaluation of the Alexa skill, the violation “there was not help specific to the user task” was only identified by the group using Nielsen’s heuristics who cited *Help and documentation* and *Recognition rather than recall*. To address the overlap and similarities between *Clarify Capabilities*, *Learnability* and *Help and documentation*, we consolidated sentences from each heuristic. We chose to remove the heuristic *Clarify Capabilities* and retitle *Learnability* to *Help and guidance*. We also moved the sentence “The system should not falsely claim to be human” from *Clarify Capabilities* to *Trustworthiness* as it relates to being truthful with the users. We added “pauses, conversation fillers, and interruptions” as examples to *Error Prevention* to address violations regarding speech recognition brought up by the Nielsen group and the experts. For example, "failed to recognize channel names" was an expert usability issue that was not found by the conversational agent heuristics.

Phase 3 Heuristics
Visibility of system status
Match between system and the real world
User control and freedom
Consistency and standards
Error prevention
Help and guidance
Flexibility and efficiency of use
Aesthetic, minimalist and engaging design
Help users recognize, diagnose and recover from errors
Context preservation
Trustworthiness

Table 3.4: The conversational agent heuristics developed in Phase 3.

3.5 Phase 4: Validation of Revised Heuristics

In the final phase, we evaluated the chatbot from Phase 3 using the revised heuristics. We found that the heuristics in Phase 3 performed well and were more suited for the voice interface, but there were needed revisions to address graphical user interfaces. In the revisions, our aim was to address violations found by Nielsen's set, but not our heuristics, for both the chatbot as well as voice to improve the heuristics' performance for both agents. After making improvements to the heuristics, we proceeded to evaluate the revised heuristics on the chatbot. We conducted online heuristic evaluations on the chatbot with 8 freelance professionals in user interface design.

3.5.1 Participants

We invited freelancers on Upwork to participate in the study. We used 'heuristic evaluation' as a keyword to filter participants and sent invitations to individuals who had above 95% job success and experience with UX/UI design. We recruited 8 participants, 4 in the Nielsen condition and 4 in the conversational agent condition, to conduct heuristic evaluations of the chatbot interface. The participants' location and experience with heuristic evaluation was balanced between the two groups. The Nielsen condition included three designers and one UI engineer. Two participants were from the United States, one from Turkey, and one from Indonesia. The conversational agent condition also included two designers, one QA test engineer, and one student. Two participants were from the United States, one from the Philippines and one from Spain. Participants were compensated between $20 to $30 depending on their hourly rate.

All of the participants had conducted heuristic evaluations between 1 to 5 times. In the Nielsen group, the participants had conducted on average 2.6 heuristic evaluations. The average self-rated level of experience with heuristic evaluation was 2.75 and experience with conversational agent design was 2.5 on a 5 point Likert scale (5 being the highest). In the conversational agent group, they had conducted on average 2.4 heuristic evaluation sessions. The average self-rated level of experience with heuristic evaluation was 3.75 and experience with conversational agent design was 2.75 on a 5 point Likert scale.

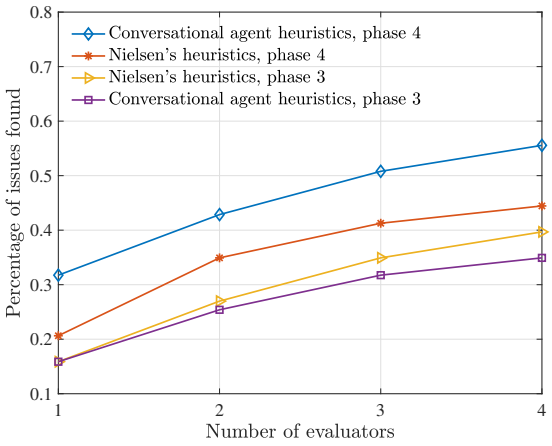


Figure 3.2: Percentage of issues for the chatbot found by the top four evaluators using the conversational agent heuristics and Nielsen's heuristics in Phase 3 and 4.

3.5.2 Results

Two of the authors iterated on the master list of usability issues for the chatbot from Phase 3 and merged in issues from Phase 4. We iterated on the master list an additional time as we found new issues that arose in Phase 4. Though the master list increased in Phase 4, we chose to compare the number of usability issues in Phase 3 and Phase 4 as they shared the same common master list. There were 63 total usability issues in the master list, including issues identified from all participants in Phase 3 and 4, and expert issues generated by the authors. Since Phase 4 had only 8 participants, we selected 8 participants from Phase 3 (the top 4 in the Nielsen group and top 4 in the conversational agent group) who had identified the most issues from the master list. To balance experience, we removed one expert participant in the conversational agent group from this selection process, who had completed more than 10 heuristic evaluations. In this analysis, we compared the 8 participants from Phase 3 and 8 participants from Phase 4. In Table 3.5, we refer to the balanced set of 8 participants in Phase 3 and 8 participants in Phase 4 as *chatbot-bal*. We refer to the set of all participants in Phase 3 and 4, 16 participants in Phase 3 and 8 participants in Phase 4, who evaluated the chatbot as *chatbot-all*. We also include the set of 12 participants from Phase 3 who evaluated the Alexa skill as *voice*.

Figure 3.2 shows that evaluators using the revised conversational agent heuristics identified more usability issues than evaluators using Nielsen's heuristics. In the conversational agent group, a single evaluator

found 20 issues, while a single evaluator found 13 issues in the Nielsen group. Four evaluators in the conversational agent group were able to find 56% of the usability issues, compared to four evaluators in the Nielsen group who found 44% of the issues.

Additionally, the final set of conversational agent heuristics performs better than the original heuristics. In Table 3.5, we see that the conversational agent group found 35 usability issues in total versus 22 usability issues found by the original conversational agent group. Interestingly, even when we consider the issues found in *chatbot-all*, we find that the four evaluators in the Phase 4 conversational agent group found more issues than the 9 evaluators in the Phase 3 Nielsen group and 7 evaluators in the Phase 3 conversational agent group (35 issues compared to 34 and 33 issues respectively).

We propose that the proportion of unique issues found by the conversational agent group is higher than those found by the Nielsen group. To test this hypothesis, we used a statistical test to compare the proportion of unique issues found by each evaluator. We consider a unique issue to be an issue found only by one heuristic set, Nielsen or conversational agent, and not found by both sets. We found that evaluators using the conversational agent heuristics found significantly more unique issues (M= 0.42, SD = 0.17), than evaluators using Nielsen's heuristics (M= 0.19, SD = 0.09), $t(6) = 2.47$, 95% CI = [-0.461,-0.002], $p<0.05$. Evaluators using Nielsen's heuristics found on average 19% unique issues.

		Phase 3		Phase 4	
Participant set	Experts	CA	Nielsen	CA	Nielsen
voice	9	30	23	–	–
chatbot-all	31	33	34	35	28
chatbot-bal	31	22	24	35	28

Table 3.5: Number of usability issues found by the experts, conversational agent (CA) and Nielsen groups in Phase 3 and 4.

We analyzed the severity of issues generated by Nielsen's heuristics versus the conversational agent heuristics. As experienced professionals in conversational agent design, four of the co-authors assigned severity ratings to the master list of issues for the chatbot. Table 3.6 illustrates the average severity rating of the issues, referred to as *severity*, and the number of severe issues (issues with a severity rating greater than 2), referred to as *num*. The overlapped group of issues found by both heuristic sets had an average severity rating of 2.5 and 2.4, in Phase 3 and 4 respectively. We found that in both phases the average severity rating of issues found only by the conversational agent heuristics is lower than issues found only

by Nielsen's heuristics. In Phase 4, the average severity rating of issues found only using the conversational agent heuristics was 1.8 compared to 2.1 for issues found only using Nielsen's heuristics. While *severity* is lower for the conversational agent heuristics, in Phase 4 the number of severe issues found is greater than Nielsen's heuristics. It should be noted that the *severity* of the overlapped issues is higher in both phases, and we suggest that the lower *severity* of the Phase 4 conversational agent heuristics is due to finding more low severity issues.

	Phase 3		Phase 4	
Heuristic set	*severity*	*num*	*severity*	*num*
CA	1.7	3	1.8	6
Nielsen	2.3	8	2.1	3
CA and Nielsen	2.5	11	2.4	18

Table 3.6: Average severity rating of chatbot issues identified only by the conversational agent (CA) group, Nielsen group, or both groups, in Phase 3 and 4.

We then grouped the usability issues to better understand the types of issues that the heuristic sets cover. The conversational agent heuristics reveal issues in the following areas.

Content

The revised heuristics address 4 out of 8 issues related to the content of the dialogue, while Nielsen's set only identified 3 of the issues in Phase 4. The conversational agent heuristics may better identify issues related to the comprehensibility of the chatbot dialgoue, such as issues with wording of questions and explanations of acronyms. There were two issues identified by the experts: "dialogue is written at an advanced reading level" and "too many chatbot messages in a row". We suggest that designers of conversational agents consider the reading level of their users.

Answer interaction

The revised heuristics address 8 out of 10 issues related to interactions with questions and responses. The conversational agent heuristics may encourage the designers to consider intuitive and free-form ways to respond to the conversational agent. Issues included users being limited to answer options that might not describe their circumstances, lack of answer validation and confusion about the "explain" feature of the

chatbot. One issue, "unclear how to submit text input", was only identified by a participant in the Nielsen group, but they did not assign it one of Nielsen's heuristics. They instead labeled it as having "no heuristic".

Guidance

The revised conversational agent heuristics identify all of the 6 usability issues sorted under help and guidance. We speculate that due to the development of the heuristic *Help and guidance*, evaluators using the conversational agent heuristics may be able to generate more issues in this area.

Humanness

The revised heuristics identified 2 out of 3 issues such as dialogue that did not appear to be genuine or engaging. One issue, "no clarification that the chatbot was not human", was identified by the original conversational agent heuristics, but not by the revised heuristics. This is likely because it was more explicitly covered in *Clarify Capabilities*. However, we think evaluators could have uncovered this issue using the *Trustworthiness* heuristic in the revised heuristics.

Data Privacy

The heuristic *Trustworthiness* was used to identify issues related to data privacy. The revised heuristics identified 2 out of 3 issues, including one issue that data was downloaded at the end of the conversation without notifying the user.

Dialogue Flow

Participants using the revised heuristics identified 5 out of 9 issues related to dialogue. The conversational agent heuristics identified many issues with the logic of the dialogue and limited control of the chatbot's topics and speed. These issues included the ordering of questions in the dialogue, the user's ability to skip questions, and incorrect utterances or follow-up questions. While the conversational agent heuristics did not identify all of the dialogue flow issues, the issues found by Nielsen's heuristics were similarly related to conversation logic and control of the dialogue.

Visual Design

The revised heuristics identified 5 of the 9 issues related to visual design, whereas Nielsen's identified 1 issue. While the conversational agent heuristics did not address all the visual design issues, these issues are generally varied and may depend on the subjective opinion of the evaluator.

Context Preservation

The original conversational agent heuristics were used to identify one issue grouped under *Context Preservation*, namely the lack of inter-session preservation. While the issue was not identified by any other participant in Phase 3 and 4, it is not a severe usability problem. Other evaluators did not record problems related to context preservation. One participant (P3) noted in their evaluation that context preservation was implemented in the interface. The chatbot interface is designed for a single interaction, and it is not intended to remember past information for multiple sessions.

The following highlight areas in which the conversational agent heuristics face limitations. There were a few issues that were largely identified by Nielsen's heuristics or by the experts.

Settings

The revised heuristics identified only 2 of the 6 issues related to the conversational agent settings. The heuristic *Help and guidance* emphasizes that guidance should be provided during the conversation. This may lead evaluators to focus less on other forms of help that exist in the interface, like the settings menu. Potential revisions could be made to address providing user guidance and feedback outside the dialogue in conversational agents with GUIs.

Audio

Both Nielsen's and the revised heuristics addressed 1 of the 5 issues regarding the chatbot's audio output. The conversational agent heuristics identified an important issue that "audio from previous messages overlaps with the current audio". The remaining issues were identified for the most part by experts, and referenced the appropriateness of using voice. We believe that *Flexibility and efficiency of use* should cover these issues raised by experts, but the heuristic may benefit from example scenarios of appropriate input/output.

Nielsen's Heuristics	Phase 4 Heuristics
Visibility of system status The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.	**Visibility of system status** The system should always keep users informed about what is going on, through appropriate feedback within reasonable time, without overwhelming the user.
Match between system and the real world The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real world conventions, making information appear in a natural and logical order.	**Match between system and the real world** The system should understand and speak the users' language—with words, phrases and concepts familiar to the user and an appropriate voice—rather than system-oriented terms or confusing terminology. Make information appear in a natural and logical order. Include dialogue elements that create a smooth conversation through openings, mid-conversation guidance, and graceful exits.
User control and freedom Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.	**User control and freedom** Users often choose system functions by mistake and will need an option to effortlessly leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.	**Consistency and standards** Users should not have to wonder whether different words, options, or actions mean the same thing. Follow platform conventions for the design of visual and interaction elements. Users should also be able to receive consistent responses even if they communicate the same function in multiple ways (and modalities). Within the interaction, the system should have a consistent voice, style of language, and personality.
Error prevention Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.	**Error prevention** Even better than good error messages is a careful design of the conversation and interface to reduce the likelihood of a problem from occurring in the first place. Be prepared for pauses, conversation fillers, and interruptions, as well as dialogue failures, deadends or sidetracks. Proactively prevent or eliminate potential error-prone conditions, and check and confirm with users before they commit an action.

Table 3.7: Nielsen's heuristics compared to the final conversational agent heuristics.

Nielsen's Heuristics	Phase 4 Heuristics
Recognition rather than recall Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.	**Help and guidance** The system should guide the user throughout the dialogue by clarifying system capabilities. Help features should be easy to retrieve and search, focused on the user's task, list concrete steps to be carried out, and not be too large. Make actions and options visible when appropriate.
Flexibility and efficiency of use Accelerators – unseen by the novice user – may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.	**Flexibility and efficiency of use** Support flexible interactions depending on the use context by providing users with the appropriate (or preferred) input and output modality and hardware. Additionally, provide accelerators, such as command abbreviations, that are unseen by novices but speed up the interactions for experts, to ensure that the system is efficient.
Aesthetic and minimalist design Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.	**Aesthetic, minimalist and engaging design** Dialogues should not contain information which is irrelevant or rarely needed. Provide interactional elements that are necessary to engage the user and fit within the goal of the system. Interfaces should support short interactions and expand on the conversation if the user chooses.
Help users recognize, diagnose and recover from errors Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.	**Help users recognize, diagnose and recover from errors** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.	
	Context preservation Maintain context preservation regarding the conversation topic intra-session, and if possible inter-session. Allow the user to reference past messages for further interactions to support implicit user expectations of conversations.
	Trustworthiness The system should convey trustworthiness by ensuring privacy of user data, and by being transparent and truthful with the user. The system should not falsely claim to be human.

Table 3.8: Nielsen's heuristics compared to the final conversational agent heuristics (cont.)

3.6 Discussion

We found that the conversational agent heuristics are useful for identifying more usability issues than Nielsen's. While usability heuristics traditionally focus on providing a clear and efficient experience, the design of conversational agent interfaces may need to go beyond usability. Providing a good user experience may require an evaluation of the conversation as well as user interactions. This is an important consideration for the integration of HCD and Implementation Science methods. Implementation scientists have proposed measuring usability as a proxy for the successful adoption of an implementation strategy [Haines et al., 2021]. Yet traditional measures of usability may not be enough to judge whether an intervention is effective for particular interfaces and contexts. Usability testing of new heuristics for conversational user interfaces may be one promising direction. In this study, participants uncovered new heuristics to consider particular user concerns about lack of confidentiality in sharing their data and wording of the dialogue that can help to consider the reading level of users. Prior work has suggested that Nielsen's heuristics are general and do not address relevant areas of specific domains [Mankoff et al., 2003][Wei and Landay, 2018][Murad et al., 2019]. In our study, two of the participants indicated that Nielsen's heuristics were not applicable to the chatbot interface. Each of these participants were among the top 4 evaluators in Phase 3 and 4 who identified the most usability issues. In Phase 3, there was one participant in the Nielsen group who created their own heuristics, titled "System Error", "Wording" and "Unexpected", for 4 of the 12 usability issues that they found. The participant brought up issues that they believed Nielsen's heuristics did not address, including: "overlapping audio", "the wording of the chatbot dialogue" and "lack of confidentiality". In Phase 4, one of the participants in the Nielsen group wrote in "no heuristic" for 3 of their 6 usability issues. In their comments, P4 said "I chose not to write [heuristics] because of confusion to categorize it." The issues labeled with "no heuristic" included: "the chatbot's utterances and questions were not applicable to their situation", and "it was not clear how to submit text input". The use of the conversational agent heuristics may have been helpful in identifying these issues. Out of the issues, we believe that there is a mapping of "lack of confidentiality" to *Trustworthiness* and "non-applicable utterances" to *Context preservation* and *Error Prevention*.

In line with Grice's maxims of relevance and quality, we introduce the heuristics *Context preservation* and *Trustworthiness* to better apply Nielsen's heuristics to conversational agents. By explicitly calling out

new design principles, evaluators consider new usability issues that may not be prioritized using Nielsen's heuristics. It is important for designers to support user expectations of context preservation [Jain et al., 2018]. Participants often noted that the chatbot seemed confused when it asked unnecessary follow-up questions. Though conversational agents may have varying levels of context handling, storing the user's recent state would help to maintain relevance in the conversation. Additionally, the conversational agent should be truthful in its interactions to encourage trustworthiness [Przegalinska et al., 2019]. The conversational agent should not mislead users about its identity, nor withhold important information about how user data will be used. In the final set of heuristics, we found that the conversational agent heuristics remained aligned with Grice's Cooperative Principles. The maxim of quantity aligns with many of the heuristics, Help and guidance, Aesthetic, minimalist and engaging design and Visibility of system status. The conversational agent heuristics recognize that while the user may require information on how to interact with the conversational agent, they should not be overwhelmed with too much information. Flexibility and efficiency of use acknowledges that the use of conversational agents may be highly context dependent. Designers and developers may consider how the conversational agent will be used and what input and output modalities, and hardware, are appropriate for those scenarios. For example, conversational agents that are used in a public context may need to provide flexibility for users to submit text input if they are not comfortable using voice.

In the final set of heuristics, we found that the conversational agent heuristics remained aligned with Grice's Cooperative Principles [Grice, 1975]. The maxim of quantity aligns with many of the heuristics, *Help and guidance*, *Aesthetic, minimalist and engaging design* and *Visibility of system status*. The conversational agent heuristics recognize that while the user may require information on how to interact with the conversational agent, they should not be overwhelmed with too much information. In particular, it may be difficult to recognize the system status and remember instructions when using a voice interface. Thus, *Help and documentation* has been removed from the heuristic set and it has been adapted, along with *Recognition rather than recall*, into *Help and guidance*. Users may need feedback and guidance throughout the conversation to better understand the status of the system, how they can search for help and what options are available to them.

We also find that the maxim of manner is supported by *Match between system and the real world*,

Consistency and standards and *Help and guidance*. The conversational agent should use language that is clear and understandable. We find that the existing text of Nielsen's heuristics fits this maxim, for example "the system should understand and speak the users' language" in *Match between system and the real world* and "users should not have to wonder whether different words, options of actions mean the same thing" in *Consistency and standards*. The conversational agent heuristics further add upon Nielsen's text to encourage smooth conversations and consistent responses.

We did not make changes to *Help users recognize, diagnose and recover from errors* as identifying and recovering from errors remains important in the design of conversational agents. We made small changes to *Visibility of system status* and *User control and freedom* to adapt them to conversational interactions. For example, in *User control and freedom*, users may need an option to "effortlessly leave the unwanted state", rather than a "clearly marked 'emergency exit'", since users may express their desire to leave the interaction in different ways, and it may be difficult to mark an "emergency exit" in a voice interface. In *Error prevention*, we expanded on the heuristic to suggest preparing for errors in conversations, as it may not be possible to eliminate all errors in dialogue based systems. Finally, *Flexibility and efficiency of use* acknowledges that the use of conversational agents may be highly context dependent. Designers and developers may consider how the conversational agent will be used and what input and output modalities, and hardware, are appropriate for those scenarios. For example, conversational agents that are used in a public context may need to provide flexibility for users to submit text input if they are not comfortable using voice.

Heuristic evaluation has been proposed as one usability testing method for identifying the design limitations of an intervention [Haines et al., 2021]. Based on these findings, heuristic evaluation is an effective approach in the design and evaluation of CUIs. We found that the usability issues identified by the conversational agent heuristics were on average lower than those found by Nielsen, as the conversational agent heuristics found more issues, and that these issues were lower in severity rating. In other words, both heuristic sets found issues similar in severity, but the conversational agent heuristics additionally resulted in more less-severe issues. The lower severity rating of these issues may be due to a number of visual design issues that were identified and assigned low priority. While it is important to identify severe usability issues, having a more complete list of usability issues, even less severe ones, can provide a better picture of a user's

experience interacting with the system. In addition, identifying an issue doesn't mean that designers have to prioritize fixing it. The same issue might be considered more or less severe depending on the target audience and context of use. Being aware of the minor issues can help designers not to exacerbate them (or introduce new similar ones) when formulating solutions to fix the prioritized issues. It is also important to consider the conversational agent that was tested in this study. The purpose of the chatbot was to collect health-related information, rather than engage the participants in purely social conversation. For example the usability issue "conversation is not engaging", identified by the conversational agent heuristics, was given a low rating, but for another type of interface this issue may be more severe and lead to short term use. These new heuristics and usability issues may be missed or not prioritized by design practitioners with expertise in heuristic evaluation.

We note the limitations to this work including the small number of participants and their level of experience with heuristic evaluation. Since participants were recruited from a large university with design programs, and from professionals on Upwork with design experience, they may have had more exposure to heuristic evaluation and UX/UI methods. Therefore, these participants may not be a representative sample of all non-experts. In the heuristic evaluation, blocked random assignment could have been used prospectively to create balance in experience levels between the groups. However, participant groups were balanced retrospectively due to the continuous recruitment of participants over three months. Additionally, while the two systems were selected to evaluate both text-based and voice-based conversational agents, there is a wide variety of conversational agent systems available that could have been used to demonstrate the effectiveness of the heuristics. We recommend that future studies evaluate how the guidelines can be applied across subject domains, usage contexts and devices.

When planning the study, COVID-19 did not influence our initial study design as Phase 3 was conducted prior to COVID-19. We designed the study to minimize participation barriers, for example the chatbot evaluation was conducted online to enable broad recruitment and the Alexa skill evaluation was in person as it required an Amazon Echo device. That said, COVID-19 did partially factor into our decision to focus on the chatbot interface in Phase 4. While it made sense for us to focus on the chatbot given our results from Phase 3, we also opted not to replicate the voice interface because of challenges with the in-person study.

3.7 Summary of Contributions to Thesis

In this chapter, we adapted usability guidelines for conversational agents to support the design and development of conversational user interfaces. Our work demonstrated that new usability heuristics can be leveraged to address unique issues to conversational systems, and improve the overall usability of conversational user interfaces.

My work contributes a set of validated heuristics that researchers and practitioners may use in their formative evaluation of conversational agents. By demonstrating their effectiveness in real world system evaluations, we propose that our heuristics can be applied to text and voice-based conversational agents. We found that four evaluators identify more usability issues when using our heuristics. These results are consistent with past work indicating that adapting Nielsen's heuristics is an effective method. We propose that the conversational agent heuristics are useful for highlighting issues related to dialogue content, interaction design, help and guidance, human characteristics, and data privacy. More broadly, our work contributes to existing research on heuristic evaluation and further highlights how this technique may be adapted for new and future interfaces.

My work in the heuristic evaluation of conversational agents demonstrated that human-centered methods - discount usability testing methods - can be applied with new heuristics to improve usability and engagement with new technologies. To improve engagement with conversational interfaces, I developed and validated usability heuristics and revealed design issues relevant to conversational user interfaces.

Chapter 4

Predictors of chatbot implementation fidelity

In this chapter, I focus on the implementation of a chatbot for social needs screening in an emergency department. There is growing interest in screening for social needs to understand and address the link between health inequities and social determinants of health (SDoH)[Marmot, 2005; Gottlieb et al., 2017; Adler et al., 2016; Marmot et al., 2008; Sulo et al., 2017], the conditions in which people are born, grow, live, work and age. Social needs are the needs of an individual as a result of their social determinants of health, such as housing instability, food insecurity or unemployment [Auerbach and Castrucci, 2019]. Hospital emergency departments (EDs) may be one appropriate place for social needs screening as EDs serve vulnerable populations with a high prevalence of social needs [Malecha et al., 2018; Gordon, 1999]. However, SDoH screening in the ED is not routine, and even when needs are identified, referral to community services and follow up may be beyond the current capacities of many EDs.

Patients benefit from assistance to complete screening and contact community resources [Hsu et al., 2020]. Yet implementing face-to-face social needs screening and referral in the ED is challenging due to anticipated patient discomfort, and clinician burden [Tong et al., 2018; Persaud, 2018]. Self-administered screening could overcome these challenges. Potential approaches include patient-facing surveys distributed via paper [Bleacher et al., 2019; Zulman et al., 2020; Power-Hays et al., 2020], automated phone calls and text messaging [Chang et al., 2022], tablets [Berger-Jenkins et al., 2019; Katz-Wise et al., 2017; Gottlieb

et al., 2014], and personal health records [Tai-Seale et al., 2019]. While patient- and provider-facing SDoH screening tools exist, such as direct entry by patients or providers into electronic health records (EHRs) [Gold et al., 2018], they face limited uptake. There are well-known disparities in patient adoption of online portals and use of personal health records [Ancker et al., 2011; Singh et al., 2022]. One reason for non-use of patient portals includes privacy and information security concerns, which indicates the importance of building patient trust in communication systems [Anthony et al., 2018]. Thus, getting patient input is important to address patients' health and social needs [Wu et al., 2019; Capp et al., 2017; Kaufman et al., 2014; Lin et al., 2017; Wilcox et al., 2018] and to design more accessible and trustworthy approaches to better engage patients. We have little knowledge about patients' sharing practices around social needs-related data during real-world clinic visits. In the context of social needs screening, conversational user interfaces (CUIs) may serve as an alternative face-to-face screening and referral, to address barriers such as anticipated patient discomfort and insufficient time, support and resources among clinicians [Tong et al., 2018; Persaud, 2018].

We have not established if patients find chatbots feasible and acceptable for social needs screening, nor whether the ED is an appropriate site for social needs screening chatbots. This is important to understand as it could lead to a screening process that is more likely to result in patients receiving care for social needs. Drawing from usability insights in Chapter 3, we applied the conversational agent heuristics to the chatbot for social needs screening. We addressed the usability problems that arose in the heuristic evaluation sessions, and applied the heuristics to ensure consistency in the interface functionality (Consistency and standards) and keep users informed about what is going on without too much information (Visibility of system status), for example. However, based on prior literature, more implementation studies of CUIs in real-world environments are needed to understand and improve the adoption of CUI-based health interventions [Kocaballi et al., 2022]. There is a lack of knowledge regarding implementation considerations for conversational user interfaces, such as how this technology might fit into practice settings, staff needs, and ongoing maintenance required. The assessment of CUI-based health interventions may also benefit from the use of standardized measures. In this work, I employed the use of implementation outcome measures and the Consolidated Framework for Implementation Research (CFIR) in the analysis of interview data. I used concurrent triangulation as a mixed methods approach to confirm findings within our study and propose

future considerations for the chatbot intervention. I investigated three implementation outcome measures [Weiner et al., 2017] to evaluate the success of a chatbot implementation for social needs screening at a large hospital ED. The use of these measures and the CFIR framework informed the development of our interview guide. This allowed us to structure conversations with ED patients around the implementation measures and processes to understand individual, contextual, and intervention-related factors surrounding the chatbot intervention. Our aim was to address the following research questions: (1) How do patients rate the acceptability, feasibility, and appropriateness of a chatbot implementation in the ED for social needs screening? (2) What are patient perceptions of using a chatbot for social needs screening? We build on prior work to evaluate a chatbot implementation situated within the ED workflow, and investigate patient perceptions of the screening and resource provision in the ED context.[1]

4.1 Methods

4.1.1 Study design

In this study, we deployed a chatbot for social needs screening in a real-world context to understand patients' perspectives on the acceptability, feasibility, and appropriateness of using the tool in the ED. We used concurrent triangulation as a mixed methods approach to confirm and corroborate findings within our study [Creswell and Creswell, 2017]. First, we collected ratings of implementation measures via surveys with participants who completed screening using the chatbot. Second, we conducted follow-up interviews with a subset of participants to further understand patient perspectives.

4.1.2 Setting and recruitment

The study took place in the ED at Harborview Medical Center, a large, public, tertiary care teaching hospital, in the Pacific Northwest region of the United States from November 9, 2020 to February 28, 2021. Patients were approached by a research assistant after completing ED registration and triage. They were considered eligible if they were at least 18 years old, English or Spanish-speaking, and did not have an acute medical or psychiatric condition. We used the Emergency Severity Index (ESI) as the qualification for identifying

[1] In this work, references to "we" refers to our work with collaborators including researchers in human-centered design and biomedical health informatics, emergency medicine physicians, and staff coordinators.

patients who would be able to participate in the study [Tanabe et al., 2004]. Patients were considered eligible if they had an ESI of 3-5 (i.e., not requiring immediate medical attention based on triage algorithm). Study procedures were approved by the University of Washington Institutional Review Board (IRB) and received a waiver of written consent. In the chatbot screening, participants read a short introduction to the study and were asked if they consent to participating by clicking "Okay, let's start" to proceed.

4.1.3 Collection of social needs and implementation measures

The chatbot for social needs screening provides relevant community resources to ED patients (Figs. 4.1 and 4.2). Participants interacted with the chatbot on an iPad and could use optional disposable headphones. The screening was available in English and Spanish. Participants used the chatbot to answer 16 questions about their social needs that were adapted from the Accountable Health Communities Health-Related Social Needs (AHC HRSN) Screening Tool [for Medicare and Services, 2019], the Benefits Eligibility Screening Tool (BEST) [Center], and the Los Angeles County Health Agency (LACHA) screening guide [Johnson et al., 2019] (see Appendix - Chapter 4: Screening Questionnaire).

At the end of the screening, the chatbot asked participants to rate three implementation outcome measures to assess the acceptability, feasibility, and appropriateness of the chatbot on a Likert scale from 1 "completely disagree" to 5 "completely agree." Using these measures, "acceptability" assesses the perception that a given innovation is agreeable or satisfactory, "appropriateness" assesses the perceived compatibility of the innovation for a given issue and practice setting, and "feasibility" assesses the extent to which the innovation can be successfully used or carried out [Weiner et al., 2017].

Participants were also asked 6 demographic questions about their age, gender, race/ethnicity, education, relationship status, and insurance status. Finally, participants were asked if they would be willing to take part in a follow-up interview. Participants were eligible for a follow-up interview if they had a working phone number. Upon completion of the screening, the participant was handed a printed copy of their responses and a list of matching community resources (Fig. 4.2), and encouraged to share their responses with their ED care team. Participants could optionally send their responses and resource list to themselves via email and text.

4.1.4 Chatbot design

HarborBot is a web application that is accessible on mobile phones and desktops. The chatbot interacts with users through chat and voice (output only) in a scripted dialogue. The front end web application is hosted on Google Cloud, developed using HTML, CSS and Javascript, and uses Python to communicate with multiple API services. Figure 4.1 shows the graphical user interface for HarborBot. We used BotUI (https://botui.org/), a Javascript framework, to build the chatbot user interface, and REDCap database [Harris, 2012] to store user responses. BotUI supports different types of interaction elements for the user to respond, including yes/no, multiple choice, select all, and free-text responses. During screening, users have the option to skip questions they do not want to answer. They may also edit previous responses by using the pen/edit icon.

After screening completion, social needs are highlighted in red and relevant resources are brought to the top of the page. Figure 4.2 shows the graphical user interface for HarborBot's summary and resource page. We compiled a list of local community resource organizations based upon resources distributed by social workers at the Harborview Medical Center ED. These resources were drawn from the Emerald City Resource Guide [Change, 2021] and Washington 211 [211, 2022], online databases that help connect people to community resources in Seattle and Washington state. The resource list included the organization name, website link, phone number, location, availability of language services, and whether the resource is free of cost.

Responses that indicate a social need are highlighted with a red icon for providers to easily view, and the corresponding resources are brought to the top of the resource list. We followed the BEST, LACHA, and AHC HRSN Screening Tool's scoring instructions on what responses constitute a social need for each domain, which then determined if the corresponding resource is highlighted on the page. In the output, all the resources were included to ensure that participants had access, regardless of whether or not they chose to disclose their social needs. After the resource list is displayed, participants can optionally send their responses and resource list via email and text which uses the Twilio API.

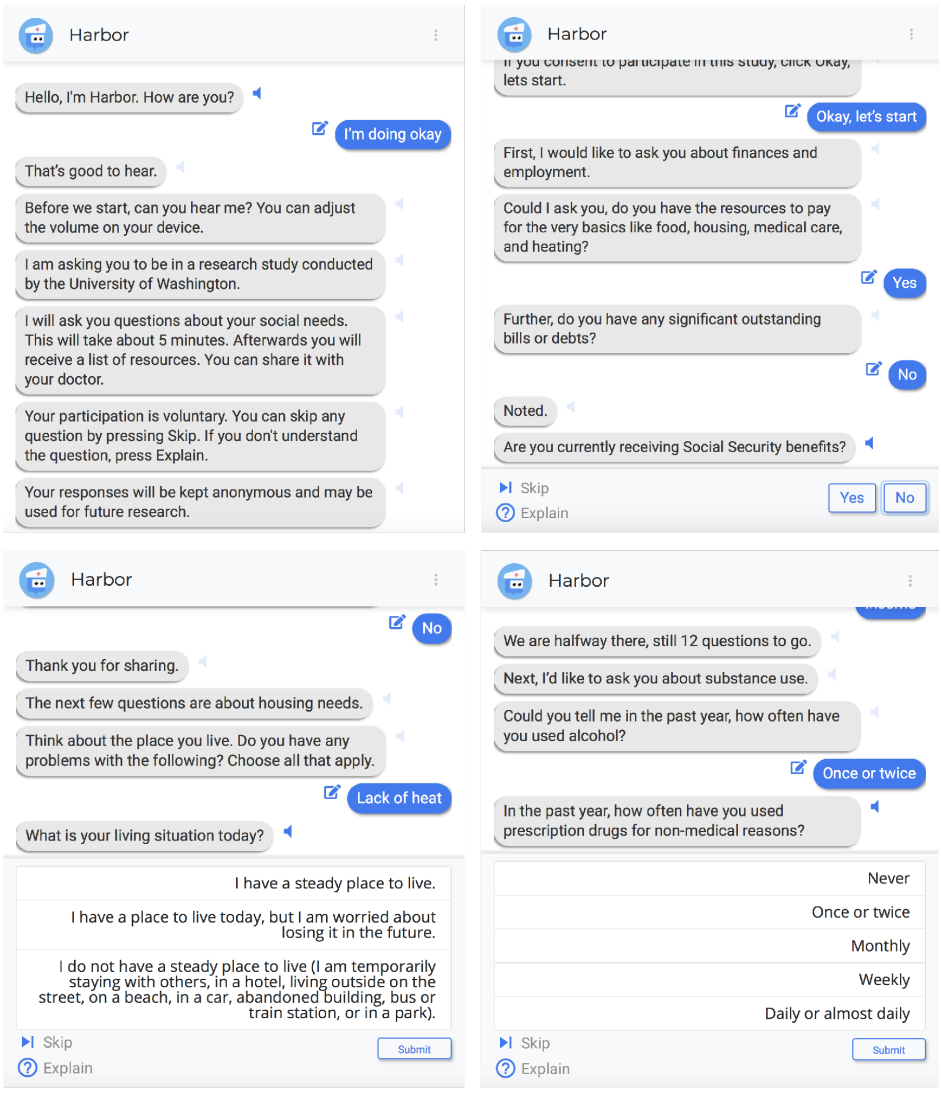


Figure 4.1: Screenshots of user interaction with HarborBot for social needs screening.

4.1.5 Follow-up interview

We interviewed participants about their experience using the chatbot. Participants were contacted via email or text message accompanied by a phone call two to four days after their ED visit. The follow up interview was either conducted at the time of contact or scheduled for a later date. The interviews were conducted by phone and were audio recorded, except for one participant who did not consent to be recorded. These interviews were semi-structured and asked participants about their perceptions of whether the chatbot was an acceptable, feasible, and appropriate way of screening (see Appendix - Chapter 4: Interview Guide). We

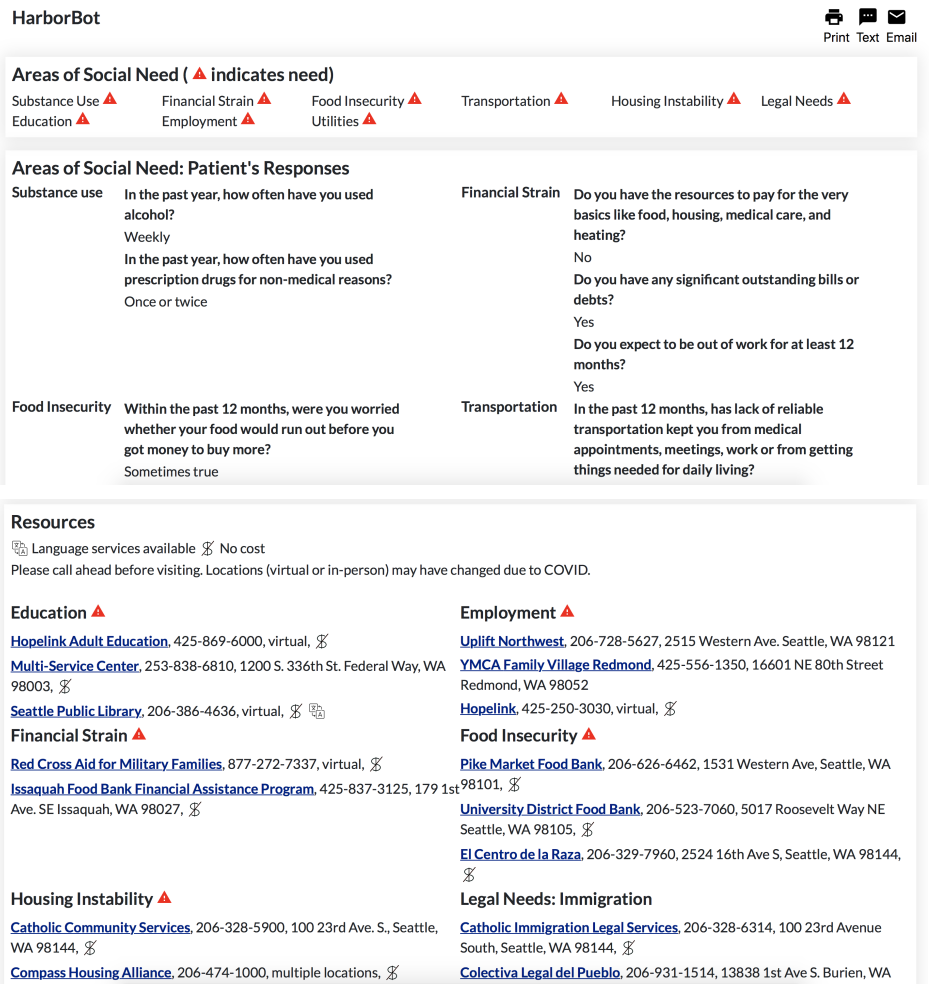


Figure 4.2: Screenshots of chatbot screening output with user responses and list of tailored community resources.

also asked participants how they used the resource list, how they currently search for and access community resources, and in what ways a chatbot could facilitate this process. The Health Literacy Single Item Literacy Screener (SILS) is a single item question that was administered to identify adults with limited reading ability [Morris et al., 2006]. Participants were offered a USD30 gift card after the interview.

4.1.6 Data analysis

We used descriptive statistics to analyze the participant demographic information (Table 4.1) and implementation ratings (Table 4.3). Analyses were performed using Microsoft Excel (version 16.43) and RStudio (version 2022.12.0+353). We followed an inductive-deductive thematic approach [Hsieh and Shannon,

2005] in the analysis of the interview data. Three team members performed inductive coding on an initial set of three interviews. Four team members then clustered the codes to develop a codebook. We incorporated concepts from the Consolidated Framework for Implementation Research (CFIR) framework [Damschroder et al., 2009] to draw from established concepts in implementation theory. Once all four team members reached agreement on the codes, we applied the codebook to the remaining interviews.

Transcript coding was divided among the four team members, and during each iteration of coding, team members coded one to two different transcripts. In research meetings, questions or concerns related to particular excerpts were discussed. Each team member reviewed the transcripts, and disagreements were discussed to achieve consensus. We returned to the initial interviews to recode them with the finalized codebook. We continued discussions across all the interviews to identify themes and patterns in the interviews to explain the ratings and provide additional insights.

4.2 Results

4.2.1 Participant characteristics

A total of 832 patients were approached and 410 patients (49%) agreed to participate in the study. Of those who agreed, 353 patients completed the screening and 3 patients under the age of 18 were removed. There were 350 participants who consented and completed the screening. The participants who completed screening ("screened participants") ranged in age from 18 to 90 years old (mean 40.7, SD=14.7) and were diverse in age, race/ethnicity, education, and insurance status, and nearly half were single or never married (Table 4.1). Among the participants, 329 participants completed the screening in English and 21 participants in Spanish. The screening took 10.92 minutes on average (SD=7.50).

Of the 350 participants, 22 agreed to follow up interviews. We conducted follow-up phone interviews and qualitative analysis concurrently until reaching thematic saturation [Hennink et al., 2017]. Interview participants (P1-P22) ranged in age from 18 to 68 years old (mean 40.6, SD=14.4). They were largely representative of the demographics in the screened participant sample, with a larger representation of White/-Caucasian participants and smaller representation of those who received some college or less. Three interview participants (13.6%) reported that they 'sometimes' need help to read written health material. The

interviews lasted on average 42 minutes.

4.2.2 RQ1: Patient ratings of the chatbot implementation: acceptability, feasibility, and appropriateness

Our findings demonstrate the value of the chatbot which was rated by participants as an acceptable, feasible, and appropriate means of social needs screening, with average ratings of 3.93 (SD=0.99), 4.20 (SD=0.86), and 4.10 (SD=0.86) respectively (Table 4.3). Figure 4.3 shows the Likert scale rating distribution for acceptability, feasibility, and appropriateness of the chatbot. The majority of participants agreed that they liked using the chatbot and it was easy to use and appropriate, with some discrepancy among the acceptability ratings (Fig. 4.3). Figures A.1 to A.3 and Tables A.3 to A.5 show the Likert rating response distribution by age, ethnicity, and education. There were some differences in perceptions of acceptability between age groups, ethnicities, and education levels. The percentage of the participants who agreed or completely agreed that the chatbot is acceptable was 88.9% among younger participants aged 18-25, compared to 65.0% among participants more than 66 years old. Additionally, 79.7% of Black, African American or African participants agreed or completely agreed that the chatbot is acceptable, compared to 53.9% and 61.5% of Asian participants and Other participants (who identified as Native American, Pacific Islander, or Middle Eastern). Participants who completed less than high school, some college or were a high school graduate, 78.1%, 79.5% and 77.3% respectively, agreed or completely agreed that the chatbot is acceptable to a greater extent than participants in graduate school or who completed some high school, 66.7% and 66.7%.

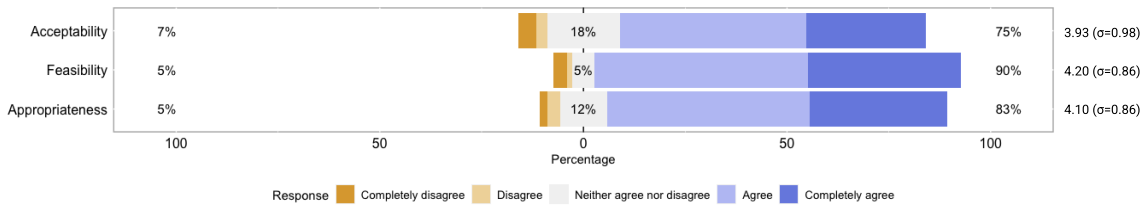


Figure 4.3: Diverging stacked bar chart of Likert scale ratings for acceptability, feasibility, and appropriateness, accompanied by mean and standard deviation for each measure. The percentage of positive responses (agree and completely agree) is stacked on the right and the percentage of negative responses (disagree and completely disagree) is stacked on the left, with neutral (neither agree nor disagree) in the center.

Table 4.1: Study participant demographics

		Screened participants (n=350), n %	**Interview participants (n=22), n %**
Age (y)	18-25	29 (8.3)	2 (9.1)
	26-35	83 (23.7)	6 (27.3)
	36-45	57 (16.3)	4 (18.2)
	46-55	36 (10.3)	3 (13.6)
	56-65	23 (6.6)	2 (9.1)
	>66	20 (5.7)	1 (4.5)
	Prefer not to answer	102 (29.1)	4 (18.2)
Gender	Male	187 (53.4)	11 (50.0)
	Female	135 (38.6)	11 (50.0)
	Additional gender category	16 (4.6)	0 (0.0)
	Prefer not to answer	12 (3.4)	0 (0.0)
Racial/Ethnic Background	White	134 (38.3)	11 (50.0)
	Black, African American or African	78 (22.3)	4 (18.2)
	Latin American, Central American, Mexican or Mexican American, Hispanic or Chicano	53 (15.1)	2 (9.1)
	More than one race	38 (10.9)	2 (9.1)
	Asian: Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other	15 (4.3)	2 (9.1)
	Other	14 (4.0)	0 (0.0)
	Prefer not to answer	18 (5.1)	1 (4.5)
Education	Some college	83 (23.7)	5 (22.7)
	High school graduate	72 (20.6)	4 (18.2)
	Bachelor's degree	38 (10.9)	6 (27.3)
	Less than high school	36 (10.3)	2 (9.1)
	Some high school	31 (8.9)	2 (9.1)
	Graduate school	31 (8.9)	1 (4.5)
	Associate degree	28 (8.0)	2 (9.1)
	Prefer not to answer	31 (8.9)	0 (0.0)
Relationship Status	Single/never married	158 (45.1)	11 (50.0)
	Married	62 (17.7)	1 (4.5)
	Divorced	47 (13.4)	7 (31.8)
	Committed relationship/partnered	30 (8.6)	3 (13.6)
	Separated	15 (4.3)	0 (0.0)
	Widowed	8 (2.3)	0 (0.0)
	Prefer not to answer	30 (8.6)	0 (0.0)
Health Insurance	Medicaid	90 (25.7)	5 (22.7)
	No health insurance	56 (16.0)	2 (9.1)
	Employer provided	53 (15.1)	3 (13.6)
	Medicare	53 (15.1)	5 (22.7)
	Don't know	28 (8.0)	4 (18.2)
	Other	24 (7.0)	2 (9.1)
	Charity Care	8 (2.3)	0 (0.0)
	Private health insurance	8 (2.3)	0 (0.0)
	COBRA	1 (0.3)	1 (4.6)
	Prefer not to answer	29 (8.3)	0 (0.0)

Table 4.2: Study participant demographics (cont.)

		Screened participants (n=350), n %	**Interview participants (n=22), n %**
Health Literacy Single Item Literacy Screener (SILS)	1 - Never	-	10 (45.5)
	2 - Rarely	-	9 (40.9)
	3 - Sometimes	-	3 (13.6)
	4 - Often	-	0 (0.0)
	5 - Always	-	0 (0.0)

Table 4.3: Ratings of implementation measures

Constructs	Implementation outcome measures	Sample size of respondents, n (%)	Median rating (IQR)	Average rating (SD)	Interview participants Average rating (SD)
Acceptability	I like the use of this chatbot to answer these questions	297 (84.9)	4 (1)	3.93 (0.98)	3.95 (1.02)
Feasibility	Using this chatbot to answer these questions seems easy to use	301 (86.0)	4 (1)	4.20 (0.86)	4.29 (0.64)
Appropriateness	Using this chatbot to answer these questions seems suitable	302 (86.3)	4 (1)	4.10 (0.86)	4.38 (0.59)

Abbreviations: IQR, interquartile range: SD, standard deviation.

4.2.3 RQ2: Patients' perceptions of using the chatbot for social needs screening

Analysis of the interviews identified 6 qualitative themes that describe ways in which participants perceived the chatbot as acceptable, feasible, and appropriate, and potential barriers to use.

Acceptability

Participants were satisfied that the chatbot provided a responsive interaction which acknowledged patients' answers and replied with personalized resources. Additionally, they liked how the chatbot afforded privacy during information disclosure, but raised questions about the security of their data. Participants appreciated the chatbot screening as an important first step in fostering a sense of care at the ED, while noting that it is important to follow-up with patients to ensure they access resources.

Chatbot provides responsive, engaging interaction Overall, participants found that the chatbot was responsive and engaged them during screening. Participants liked that the chatbot maintained the responsiveness of a human interaction and guided them through each question.

If you put yes or no, depending on what you put, you have another answer on the chatbot. If [the screening] was just a couple of questions on paper, then you wouldn't receive that reply. (P17)

Participants also liked that the chatbot provided personalized recommendations for community resources, avoiding information overload through extraneous recommendations. They appreciated that the conversation was brief, rather than repetitive, unlike past surveys that asked many similar questions about the same type of social need. P13 was looking for food assistance and found that the resources were tailored to their social needs.

It seems more personal because it literally narrows down and takes out what you said yes to, what you said no to. Then it only gives you information on what you need help with, instead of giving you a load of information on certain things that [are not relevant to you]...say, if you're not an alcoholic, it's not giving you a number to AA...If you need a food bank, it's giving you a number to food banks, it's giving you a number to donation places. (P13)

Chatbot helped preserve privacy during information disclosure, but prompted questions about data sharing and security Participants who did not want to be overheard in the ED valued the chatbot. They liked that they could input their responses instead of speaking out loud.

I'd rather answer my questions and everything with the chatbot. That way [other patients are] not hearing what's going on with me as far as money. In fact, a lot of the things that I don't like is I have to repeat, like give them my address, my phone number, in front of these strangers who you don't know...I don't want to give out that kind of information out loud...since we were online and the chatbot actually submitted the information that I sent out, that was actually probably one of the things that I felt safest with at that time. (P6)

There was a sense that the ED was not a secure place to discuss personal information and the chatbot afforded privacy from answering questions in an open space. P6 was not only worried about being overheard, but worried about other ED visitors who might take and view their responses if they were on paper.

There's a couple times where I was in the ED by myself. I've been medicated with morphine or something, and I have been out there and waiting for a cab, and this person would sit next to me to try to grab whatever was in my bag. (P6)

Privacy during information disclosure was very important to participants to avoid direct judgment or stolen information. Participants desired that their information be stored securely in the EHR after the chatbot interaction, and assumed that their information would not be shared with unauthorized individuals. However, some participants were cautious of what information to share with the chatbot as they felt it may lead to stolen information. P10 was hesitant about sharing personal information via the chatbot and explained that they try to be careful no matter what application they use.

I have to try to be careful what app I'm using or whatever... because there are predators out there that will steal your identity. (P10)

Together, these examples illustrate that participants found privacy-preserving aspects of the chatbot to be acceptable, including no requirement to speak responses out loud, and assurances that responses would not be shared inappropriately. However, data security was a concern that reduced acceptability.

Feasibility

Participants found that the chatbot was a feasible method of social needs screening in the ED. They found the chatbot easy to use, understand and quick to complete.

Chatbot is easy to use and understand Participants found the chatbot easy to use which facilitated the successful completion of screening. In support of their high ratings of feasibility, participants said they could easily understand and answer the questions.

The instructions were pretty self-explanatory. You could understand the instructions, like when it was dragging you to the next page and what to do and all that. So that was pretty cool that they broke everything down for you as you went along...I didn't have to ask [the research assistant] anything the whole time I did it. (P5)

P13 agreed the chatbot was easy to use and compared the experience to playing a computer game. Further, P5 liked using the tablet and selecting multiple choice options rather than typing because their hand was broken. P10 described themselves as less familiar with technology, but still found the chatbot as easy to use: "I don't dislike it, but I'm just used to doing regular straight paper, not a tablet. I'm not there yet…I'm not knowledgeable like some other people." (P10)

Chatbot screening is quick to complete When asked about how easy the chatbot was to use, participants found the chatbot feasible because it could be used quickly and easily. The screening did not take a lot of time to complete: "It was faster...more convenient maybe than talking to the representative directly" (P3). The chatbot was direct and easy to understand, whereas people may not be as direct: "You just answer Yes or No, it's not that difficult" (P2). P16 thought it was an efficient and effective way to get responses since they had free time in the ED waiting room and they would not be motivated to complete a survey sent via email.

> *I think you have some free time…[compared to] a survey that comes through email, I know I get them all the time and almost never filled them out. So the way in which the survey is administered [via chatbot], I think it's a good way to get more responses. (P16)*

Overall, participants reported that they did not mind filling out questions to pass the time and the chatbot only took a short time to complete.

Appropriateness

Participants perceived the chatbot as an appropriate technology for the setting. Participants were comfortable sharing their social needs with the chatbot to avoid attention from other ED visitors and social judgment present in face-to-face screening.

P1 found that the ED was busy and the chatbot was compatible with this context.

> *I think in the setting of the hospital that it was easier to use the chatbot than it would be to find a quiet place to sit down where you could have a discussion with a person. (P1)*

Most participants did not feel comfortable calling attention to themselves in the ED, and using the chatbot on the tablet seemed like a casual, normal activity that everyone was participating in. Participants cited fear of social judgment as a reason that they preferred using the chatbot: "You might open up to a chatbot and not a person... [there is] a lot of shame involved in some issues" (P4). P4 was searching for stabilized housing options and had spent the last 15 years learning about homelessness. Interacting with a chatbot has the potential to minimize social judgment that would occur if talking with a healthcare worker "because you don't have to deal with its [the chatbot's] attitude" (P6).

Participants also had different levels of comfort with what information to share with healthcare providers. They may be uncomfortable or embarrassed to talk with a healthcare provider about social needs, especially a provider they do not know. While P17 discussed how healthcare providers can be helpful to provide information about social needs and redirect them to resources, they were not comfortable with bringing up their social needs to their provider.

> *I don't feel comfortable talking about my financial situation with my doctor...They can be helpful if they provide you the information there, or they redirect you... Usually there's no conversation to bring it up... Well, they're just telling you what to do to make it better, or they're going to prescribe you something. So sometimes that conversation doesn't go along with the housing. (P17)*

P21 even hesitated to disclose information, such as their ability to pay for utilities, via the chatbot as they felt it may change the care they receive from ED providers. Others discussed receiving lower quality care at the ED based on their social needs in the past and did not want that to reoccur.

> *I don't want to tell them that I'm homeless because I feel like I'm being treated differently as opposed if I just tell them, oh, okay, well I live over here... The whole issue, I think just came down to, they found out I was homeless. I was sleeping outside. The doctor expressed that they didn't want to do the surgery because I didn't have a sterile place to heal. I said, well, that's what you guys are here for. You have respite beds that you provide for people that need a place to heal. And so the answer that I got was, well, we can't reserve respite beds. (P5)*

Screening is the first step in fostering a sense of care

The chatbot was perceived as a valuable first step in learning about social resources. P11 was homeless on and off for over 20 years and explained that screening for social needs was important because "a lot of people don't know where to look. . . [and] don't have access to the internet, so I think the way it [chatbot] was brought to me [on a tablet] in the hospital was an awesome thing." Even for those who know where to look, using the chatbot was seen as another way of accessing information, particularly since the current resources they are aware of may not be meeting their needs.

> *I just feel like the more access and the more ways of making people get the resources the better. I don't feel like there should just be one way of getting resources out to people. . . Considering a lot of the day services will give you booklets with resources. But the problem with that was the resources wouldn't be updated, so a lot of it was outdated. A lot of places you would call were closed down. They wasn't operating no more. So the booklet was useless at the end of the day. (P5)*

All participants said they would use the chatbot in the future and most were open to tools that helped them discover resources.

However, effective follow-up on patients' social needs is necessary for patients to feel cared for in the ED context. Participants mentioned that the screening should feel personal and serve a purpose beyond collecting information. P19 felt the chatbot didn't provide personal benefits: "It was just a way of filling out the survey...It didn't benefit anything really." P19 wanted to have a person in the loop to ensure that they are going to receive help.

> *It's good to have an actual human being there...telling me that they want to get you help or they can get you assistance, and then they stay there and you answer the questions that they're asking you. It feels like a more believable situation...it would be nice to maybe have someone contact you the day after you get out or a couple of days after you get out and go over what you filled out instead of just an automated voice. (P19)*

Further, patients may want to elaborate on specific answers to ensure that they get help.

It was easy to answer because it had preloaded answers... but, [you] can't elaborate too much with a chatbot... It asked, 'Are you or somebody in your household experiencing hardship?' Then, I said yes. Then, it asks how, and I said income, but I wasn't able to type in more than income... I wanted to say it was his income, not mine. That's the issue right now because we're sitting here waiting for unemployment. (P6)

For example, one participant tried to hand off the printed output to their provider, but kept being redirected to the next staff person until they were able to share their printed screening results with a social worker.

Participants rarely brought the printed responses and resource list to start a conversation with their provider. Some participants were recurring patients who felt that ED providers are very busy and did not want to bother them by bringing up their social needs. Although few participants expressed concerns about sharing social needs through a chatbot in the ED, the above mentioned concerns and preferences around sharing social needs might hinder some patients' sharing and early engagement with providers. To increase appropriateness of a chatbot for social needs screening in an ED context, patients require secure and reliable pathways for following up on resources.

4.3 Discussion

Our findings indicate that the chatbot implementation at the ED was perceived by patients as a feasible, acceptable, and appropriate form of outreach that could increase uptake. The ED has an explicit mission statement to care for vulnerable populations, and participants recognized the ED as a place where many individuals with social needs go for assistance and could participate in the screening. Those who may be more in need of resources, such as those who have not completed an advanced degree, may be more receptive to the chatbot screening, for example patients who completed less than high school may find the chatbot more acceptable than patients who completed graduate school. The qualitative responses supported the survey responses when triangulating on the data. This is significant as it suggests that chatbots could facilitate a screening process that ultimately connects patients to care for social needs, supporting the mission of EDs as part of the social safety net and improving health and well-being for members of the most vulnerable

patient populations. Providers could use social needs information to better personalize treatment plans and direct patients to resources available in the hospital and community.

However, not all participants were positive about chatbots and strategies to improve uptake in this group will be important future work. Among the interview participants, 6 out of 22 participants rated the chatbot acceptability as 3 or less on a 5-point Likert scale. This allowed us to gain insight into limitations from participants who were hesitant about using the chatbot. Those who did not want to use the chatbot described themselves as being less familiar with new technology and applications. The presence of a trained professional in the hospital ED can help to support the screening process, in particular for older patients who may find a chatbot screening less acceptable than younger patients. Some participants felt uncomfortable sharing social needs with providers in the ED after completing the screening. This was due to patients' perceived prioritization of medical needs over social needs at the ED, and the potential negative impact on their emergency care. Although prior work indicated that young patients want help with social needs from providers [Chang et al., 2022], most interview participants did not discuss their screening results with ED providers. For those who have data security concerns or do not want to discuss social needs with their providers, future chatbot design should inform patients how their data will be accessed for clinical purposes. If desired, they should be allowed to opt out of data sharing. For patients who want to elaborate on their answers, they should be provided flexibility within the chatbot interaction to express themselves and emphasize what resource they need the most assistance with.

Many of the 22 interview participants were facing social needs; 13 participants could not pay for basic needs or had significant bills, and 6 participants were facing housing instability. Some participants wanted reliable and actionable support in accessing resources, thus one future direction is to link chatbots with existing healthcare systems to facilitate referrals. It is important to establish pathways to alert providers to acute social needs, get patients in touch with community-based organizations for resource referral, and help providers follow-up on patients afterwards. The design of a chatbot for social needs screening may benefit from standardization since conversational user interfaces in healthcare can lead to unintended consequences, such as miscommunication due to information overload [Ash et al., 2004]. In the next steps, we plan to craft recommendations for system-wide implementation of the screening and referral process developed. Further, departmental and health system stakeholders plan to integrate social needs screening with existing

technologies, such as EHRs. There is ongoing research to prepopulate social needs by extracting social needs related information from clinical notes to address challenges of patient data collection [Lybarger et al., 2021].

The chatbot screening has the potential to reduce ED provider and social worker burden through EHR integration to summarize patients' acute social needs and automatic referral to the relevant department. Providers may not discuss social needs with patients because there is not an established pathway to address them. The chatbot screening can therefore help to identify and address social needs that may go unaddressed during patient visits. Without knowledge of patients' social needs, such as their inability to afford prescribed medication, the effectiveness of healthcare can be diminished. Given that patients may be concerned about social needs disclosure, health systems should facilitate social needs screening to protect patient privacy and improve treatment. One strength of the work was the insights we learned from talking with participants facing social needs, as the hope is that they would be the intended audience. The study results may have varied if it were conducted in a primary care setting as participants without health insurance coverage may be less likely to present in primary care.

There are several limitations in this study. First, our findings are largely based on participants' screening responses and interviews with a convenience sample. While we aimed to recruit participants representative of the ED patient population, self-selection bias may be present in participants who opted to participate in the study. For instance, participants who had particularly negative experiences in the ED may be less prone to participate or adopt chatbots. Second, the presence of the research team during recruitment and novelty effect of the chatbot could also have influenced their use and feedback on the chatbot. Finally, our study was conducted in a large public hospital in one geographic region of the United States, which may limit the generalizability of our findings. Despite these limitations, our study has a number of strengths, including its reach and mixed methods approach that provide important groundwork to guide future studies.

4.4 Summary of Contributions to Thesis

In this chapter, we aimed to understand the successes and challenges of social needs screening via chatbots to inform the design, development, and deployment of screening interventions. Our work demonstrated that conversational user interfaces can be leveraged to support patient needs due to its perception as a preferred,

appropriate screening modality.

We found that patients generally welcomed the chatbot use for disclosing social needs and learning about resources, but had concerns about data privacy and support in following up on resources. We also found that the ED context was conducive to social needs screening, while also facing limitations. By screening in the ED, the intervention has the potential to reach many individuals facing social needs. In some interview conversations, participants mentioned their reasons for visiting the ED, including non-medical issues, such as medical bill assistance and medication refill. Further, some participants left the ED waiting room before being admitted, due to long wait times. Thus, screening in the ED waiting room prior to admission may have a wider reach and be completed by more individuals than are actually admitted. While the chatbot-specific intervention and ED context supports patient engagement with social needs screening, individual and contextual factors limit this engagement. The emergency department may not facilitate disclosure of social needs due to perceived prioritization of medical needs, perceived negative impact on medical care, and concerns about taking time from providers. Health interventions that have been proven to improve health outcomes are typically longitudinal, tailored interventions that connect patients with community health workers (CHWs) for case management [Butler et al., 2020]. Though more institutional support is needed to follow-up with patients, chatbots may serve as comfortable first touchpoint in the patient's journey through the ED to disclose social needs.

My work in the social needs screening context shows that human-centered methods - design and interview study - and implementation science measures and methods - deployment study - can reveal implementation considerations that arise when new technology systems, such as chatbots, are introduced to the emergency care context. Towards demonstrating that chatbots are an acceptable, feasible and appropriate form of screening, I leveraged implementation outcome measures and an interview study to reveal individual, contextual, and intervention-related factors related to the implementation. We use a mixed methods approach to document ways in which context drives implementation, and provide insights for future research on CUIs for social needs screening. When context is taken into account in research, study findings are more applicable to different populations, settings, and time periods [Brownson et al., 2022]. In implementation research, a major reason for using mixed methods is to 1) measure intervention and/or implementation outcomes with quantitative methods, and 2) understand process with qualitative methods [Palinkas et al., 2011].

Qualitative methods are especially suited to understanding the complexity and variation within real world settings [Patton, 2014]. To improve engagement in social needs screening, I deployed the system in a real clinical setting and revealed opportunities for adaptation in the implementation process.

Chapter 5

Integrating human-centered design and implementation science for chatbot design

In this chapter, I discuss the integration of human-centered design and implementation science methods towards establishing an evidence base for conversational user interface (CUI) design in breast cancer screening outreach for minority women. I draw from methods across the two disciplines to understand which design aspects of CUIs may lead to improved outcomes in breast cancer screening. Breast cancer is the most common cancer among Black/African American women living in the United States (127 cases per 100,000 women), and is the second leading cause of death among Black women [DeSantis et al., 2019; Giaquinto et al., 2022; CDC, 2022]. Breast cancer screening can lead to earlier diagnosis and improved outcomes; however, racial disparities in screening mammography exist due to health inequities, such as less access to timely and high-quality prevention, early detection, and treatment services, and later stage at diagnosis [Newman and Kaljee, 2017; Daly and Olopade, 2015]. Despite health system recognition and outreach, disparities in breast cancer mortality among minority women continue to persist [Hardy and Du, 2021; Eley et al., 1994]. Prior research has demonstrated that Black women experience multiple barriers to breast cancer screening including reduced access to care, mistrust, fear of diagnosis, prior negative health care experiences, and lack of information regarding breast cancer risk [Passmore et al., 2017; Ko et al., 2020; Jones et al., 2014; Katapodi et al., 2010; Thompson et al., 2004; Orji et al., 2020; Adegboyega et al., 2019; Young et al., 2011; Molina et al., 2015]. Black women who have not been screened have limited knowledge

about mammography screening [Adegboyega et al., 2019]. Further, among women who have been screened, they have reported being inadequately informed or prepared for what to expect during the procedure. More awareness and patient education about breast cancer is needed to achieve equity in screening [Ferreira et al., 2021; Adegboyega et al., 2019; Huq et al., 2022]. Outreach and awareness is important as women are often caretakers and a source of social support for others, but they may have difficulty expressing their own need for support [Kim et al., 2018]. Fear and medical mistrust are also barriers to regular mammography screening [Adegboyega et al., 2019]. Middle-aged African American women who do not trust physicians are significantly less likely to get breast cancer screenings [Guo et al., 2019]. Trustworthiness is an important consideration for health interventions [Veinot et al., 2013] and prior research has drawn from various theories of trust to design for user interfaces [Veinot et al., 2013; Cassell and Bickmore, 2000; Benbasat and Wang, 2005]. One approach is to position research efforts in collaboration with people and institutions that users already trust, such as community-based organizations [Veinot et al., 2013]. In addition to addressing the above barriers to breast cancer screening disparities, many have advocated for earlier and more frequent breast cancer screening to equitably care for the health needs of Black women [Oppong et al., 2021; Ahmed et al., 2017; Smith-Bindman et al., 2006]. Initiating biennial screening in Black women at age 40 can reduce breast cancer mortality disparities [Chapman et al., 2021].

Tailored interventions have had some success in improving breast cancer screening rates. Past work has explored tailoring communication with users based on personality traits [Zhou et al., 2019; Bickmore and Cassell, 2001], and literacy level [Ancker et al., 2016] to improve patient uptake and comprehension. Cultural tailoring may have a positive impact on health outcomes [O'Leary et al., 2020], and prior HCI research has established the importance of making aspects of an interface culturally relevant, such as text, images, colors, and modes of interaction [Kim et al., 2020; Mendu et al., 2018; Harrington et al., 2022]. A meta-analysis of 14 randomized control trials evaluating breast cancer screening interventions for Black women demonstrated an overall modest improvement in screening rates [Copeland et al., 2018]. Although no patient or study characteristics significantly moderated screening efficacy, the most effective interventions were those that provided tailored information. Speaking to participants' religion and racial pride were among the suggestions for culturally relevant tailoring strategies. Prior research has discussed a number of facilitators to breast cancer screening, including access to health insurance and healthcare providers, older

age, higher education level, ideal patient-provider communication, personal diagnosis of cancer, having relatives and friends with diagnosis of cancer, [Agrawal et al., 2021; Davis, 2021; Guo et al., 2019]. One study suggested that interventions might focus on broadening health insurance coverage and working to improve patient-provider communication [Agrawal et al., 2021]. Prior research has evaluated breast cancer screening interventions that employed community health workers or health educators to facilitate patient navigation [Blumenthal and Alema-Mensah, 1997; West et al., 2004; Russell et al., 2010; Marshall et al., 2016; Zhu et al., 2002]. There has also been exploration of interventions using telephone and mail reminders and mailed informational materials such as letters, DVDs, and magazines [Goel et al., 2008; Hendren et al., 2014; Jibaja-Weiss et al., 2003; Gathirua-Mwangi et al., 2016; Kreuter et al., 2005]. One study found greater improvement in breast cancer screening rates using interactive computer technology compared to video or pamphlet education [Champion et al., 2006]. In this study, it was suggested that the interactive computer program was more effective because it required more concentration from participants. Yet, there is not an established evidence based surrounding the effect of more interactive interventions on breast cancer screening rates. Current evidence on breast cancer screening interventions is limited to a small number of trials and the optimal design of a tailored intervention is still unknown.

In this chapter, we discuss the design of a chatbot prototype for breast cancer screening outreach to support Black/African American women in scheduling mammograms.[1] In late 2020, we brought together a team of researchers and health system leaders at UW Medicine and the Cierra Sisters to address inequities in breast cancer screening through the design of a chatbot that could facilitate outreach. The Cierra Sisters is a local organization whose mission is to increase knowledge about breast cancer in the African-American and underserved communities. Breast cancer screening rates in the health system at the time were 61.5% among Black women compared to 73.3% among white women (internal health system data). To facilitate outreach, text messaging or the use of chatbots may be a useful and interactive modality that can reach a large number of users to improve awareness and health equity in screening [Kim et al., 2022]. Although evidence-based toolkits that encourage screening for Black/African American women exist [Komen, 2022], they have yet to be applied with a more interactive, dialogue-based modality, such as CUIs, and studied as implementation strategies. We have a limited understanding about which design aspects of conversational

[1] In this work, references to "we" refers to our work with collaborators including two human-centered design researchers, a primary care physician and early-stage investigator, a community-based organization leader, and a design team.

user interfaces lead to effective outcomes in breast cancer screening, and generally in health contexts. By integrating human-centered design and implementation science methods, we may receive more nuanced feedback about the chatbot design and which components to test, towards establishing an evidence base for CUI design in this health context.

Guided by the Multiphase Optimization Strategy (MOST) framework, we integrate human-centered design and implementation science methods to design the chatbot. MOST is a method that involves three phases (Preparation, Optimization, and Evaluation) for systematically building and evaluating interventions to ensure they comprise active components [Collins et al., 2005; Broder-Fingert et al., 2019]. Drawing from the MOST Preparation phase and applying human-centered design methods, we engaged in an Exploratory phase to (1) center stakeholders and community partners in a qualitative analysis to understand breast cancer screening determinants, and (2) to develop a conceptual model for the intervention, identify core components, and determine what outcomes should be optimized. Our team included a human-centered design researchers, a primary care physician and early-stage investigator, and a community-based organization leader. The team received project mentorship from the Optimizing Implementation in Cancer Control (OPTICC) team that includes experts in implementation science. We also included health system stakeholders in design meetings and research activities, such as health care equity leadership, primary care and population health leadership, and primary care health navigators. In this exploratory phase, we engaged prospective users in human-centered design to understand key barriers and facilitators to breast cancer screening and evaluate low and medium fidelity design prototypes in interviews and focus groups. We also applied the conversational agent heuristics to the chatbot design in Chapter 5. One heuristic that guided the design was Trustworthiness, which advises that the system should convey trustworthiness by being transparent and truthful with the user (in other words the CUI should not falsely claim to be human). After designing an initial chatbot prototype and developing a causal pathway diagram, we engaged in the MOST Optimization phase. We conducted a randomized factorial experiment of specific components and reviewed results of the experiment. In the optimization phase, we prioritized chatbot characteristics based on user input to optimize trust and engagement. We then conducted a factorial design experiment to understand which chatbot design elements impact trust, engagement, and future intention to use. Throughout this process, we iterated on the chatbot design with a interdisciplinary design team. We held regular design team meetings and iterated on

the chatbot design with stakeholders and community members. In this work, our research questions were: 1) What are determinants (e.g. barriers and facilitators) to breast cancer screening among Black women? 2) What are the perceptions of a chatbot for breast cancer screening outreach? These two research questions informed the next question in the optimization phase: 3) What is the effect of chatbot persona and messaging on trust and engagement? In our work, we integrate human-centered and implementation science methods towards an improved understanding of key determinants to breast cancer screening and guidance in the design of a chatbot intervention as an implementation strategy to improve trust and engagement.

5.1 Exploratory Phase

Key determinants (i.e., barriers and facilitators) are often identified through evidence review, and qualitative data can also be valuable in informing key determinants. The use of human-centered design methods can augment identification of key determinants and other components in Causal Pathway Diagrams (CPDs) via mockups and/or early prototypes to elicit feedback on initial design and use. The chatbot implementation strategy was prioritized as an intervention among interdisciplinary team members because of its innovation and the low resource burden in primary care with better potential for sustainability. The study protocol was reviewed and determined exempt by the University of Washington Institutional Review Board. The use of qualitative methods, such as interviews, is particularly useful because determinants can be elicited in the context of the implementation strategy – which may help to optimize determinant-strategy matching. We identified and prioritized key determinants through rapid evidence review of breast cancer screening determinants among Black women and human centered design (HCD) methods. HCD methods included semi-structured interviews including a chatbot mockup, and focus groups with end-users who were shown an early prototype of the chatbot which was iterated based on qualitative data analysis of the interviews. Interviews and focus groups were facilitated by our community engagement expert to create space for participants to provide honest and thorough feedback.

5.1.1 Methods

Rapid Evidence Review

The objective was to identify determinants to breast cancer screening among Black women emergent from recent literature. Rapid evidence review was conducted following established methods described in the National Collaborating Centre for Methods and Tools Rapid Review Guidebook [Dobbins, 2017]. We defined a research question – "among Black women in the United States, what are determinants (i.e., facilitators and barriers) to breast cancer screening?", searched for research evidence, critically appraised information sources, and synthesized evidence.

Our search strategy prioritized evidence in the past 3 years and included search terms in or related to the research question: (Mammogram, Mammography, Cancer Screening, Breast Cancer Screening), (Breast Cancer), (Women), (Black, African American, African American, Minority), (Race, Ethnicity), (Disparities, Determinants), (Facilitators, Barriers). Searches were conducted in PubMed, Health Evidence, Public Health +, and the National Institute of Health and Care Excellence. We considered studies done in the United States as the experience and impacts of individual and systemic racism differ across countries. We focused on results among Black/African American individuals given the research question and aim to identify specific determinants within this group, however we did include studies with multiple racial groups represented. We focused on studies that included individuals aged 40-74 years to match the population eligible for average-risk breast cancer screening. Publications in the 3 years prior to evidence review were prioritized acknowledging determinants may change over time (e.g., with technology advancements such as online scheduling or with allowing for mammogram scheduling without PCP referral) and in keeping with methods in the National Collaborating Centre for Methods and Tools Rapid Review Guidebook [Dobbins, 2017]. Critical appraisal was guided by the 6S Pyramid framework developed and made available by the National Collaborating Centre for Methods and Tools [NCCMT, 2023]. Data was categorized by source (i.e., search engine), study type (e.g., single study, meta-analysis), population, and results.

Interviews

The objective of the interviews was to elicit determinants to breast cancer screening among Black women living in western Washington as well as feedback about an initial mockup of the chatbot. The interview

guide was developed by our research team with additional input from members of the Breast Health Equity committee – a health system committee including operational leaders, physicians, and researchers dedicated to addressing disparities in care related to breast cancer screening, diagnosis, and treatment (see Appendix - Chapter 5: Interview Guide). Questions focused on determinants to breast cancer screening and past experiences with breast cancer screening. Additionally, two members of the research team created a mockup of the chatbot tool including several mockups of a chatbot for breast cancer screening outreach. We used convenience sampling through fliers posted in primary care clinics and email to the research team's established community networks to identify and recruit individuals who identified as Black women between the ages of 40 and 74 years and lived in either King or Pierce counties in Washington state.

We recruited 21 individuals which we estimated would be sufficient to reach thematic saturation [Hennink et al., 2017]. All interviews (n=21) were conducted by two members of the research team; the community engagement lead on the team conducted the vast majority of interviews (n=18). Interviews were conducted via Zoom videoconferencing technology, audio recorded and transcribed. In addition to questions regarding determinants to and experience of breast cancer screening, participants were shown screenshots of the initial mockup for the chatbot tool and asked specific questions for feedback (Fig. 5.1). Four members of the research team read and coded the transcripts to generate and refine themes through several iterations until consensus was reached. Each interview was analyzed and coded once by individuals on the research team using a directed content analysis approach with both deductive and inductive analysis [Hsieh and Shannon, 2005]; codes were then discussed as a team. Deductive codes were created using prior research organizing breast cancer screening barriers as personal, structural, and clinical [Young et al., 2011]. Inductive codes emerged from a close reading of an initial subset of the transcripts and were added to the codebook. Qualitative data analysis resulted in themes around the chatbot design, and barriers and facilitators of breast cancer screening. We facilitated an ideation workshop with the research team and Breast Health Equity committee to brainstorm how this research might address the themes brought up in the interviews. We used a 2x2 prioritization matrix as a tool to identify the most impactful and feasible ideas that arose. This analysis was used to develop an early chatbot prototype.

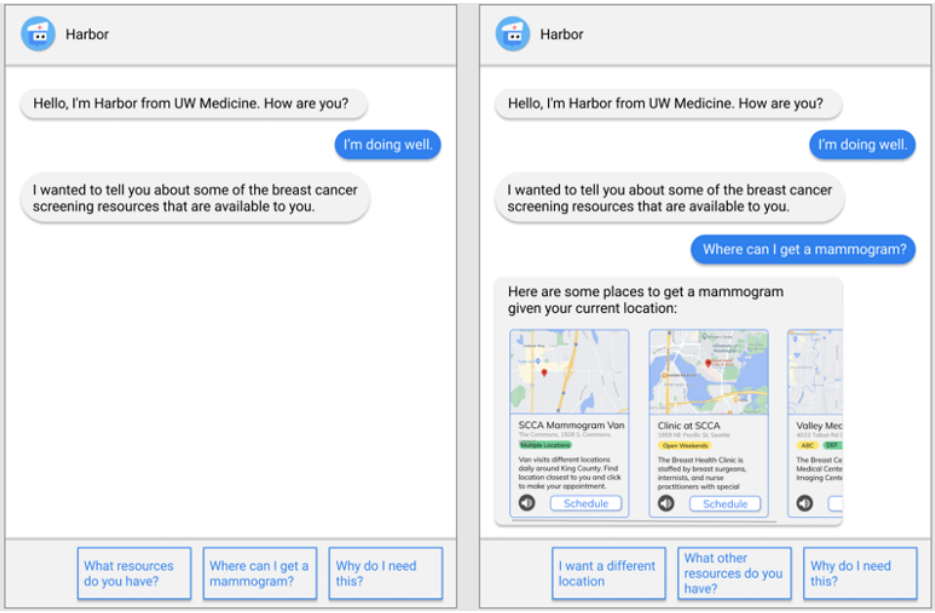


Figure 5.1: Initial mockup of the chatbot tool.

Focus Groups

The objectives were to elicit feedback on an early static prototype of the chatbot tool informed by the interviews. The research team developed an early static prototype of the chatbot tool iterating on the initial mockup using themes and feedback that emerged from qualitative data analysis of the individual interviews (Fig. 5.2). In addition to the prototype screens, short videos were included with questions and answers to questions such as – "Why should I get screened?", "What can I expect from a mammogram?", "What happens if the mammogram is abnormal?". The interview guide for the focus groups was developed by our research team with additional input from members of the Breast Health Equity committee (see Appendix - Chapter 5: Focus Group Guide). The guide included questions about perceptions of, engagement with, and usability of the chatbot based on the prototype screens and videos.

The same convenience sampling methods were used for the focus groups as were used for the individual interviews. We conducted 3 focus groups with a total of 9 participants. Focus groups were led by the community engagement lead and joined by multiple members of the research team. Participants were shown three example interactions with the chatbot prototype, 1) patient-initiated scheduling of a mammogram, 2) system-initiated patient education, and 3) system-initiated re-scheduling, and asked specific questions for feedback. The same procedures were followed as for the individual interviews. We used template analysis

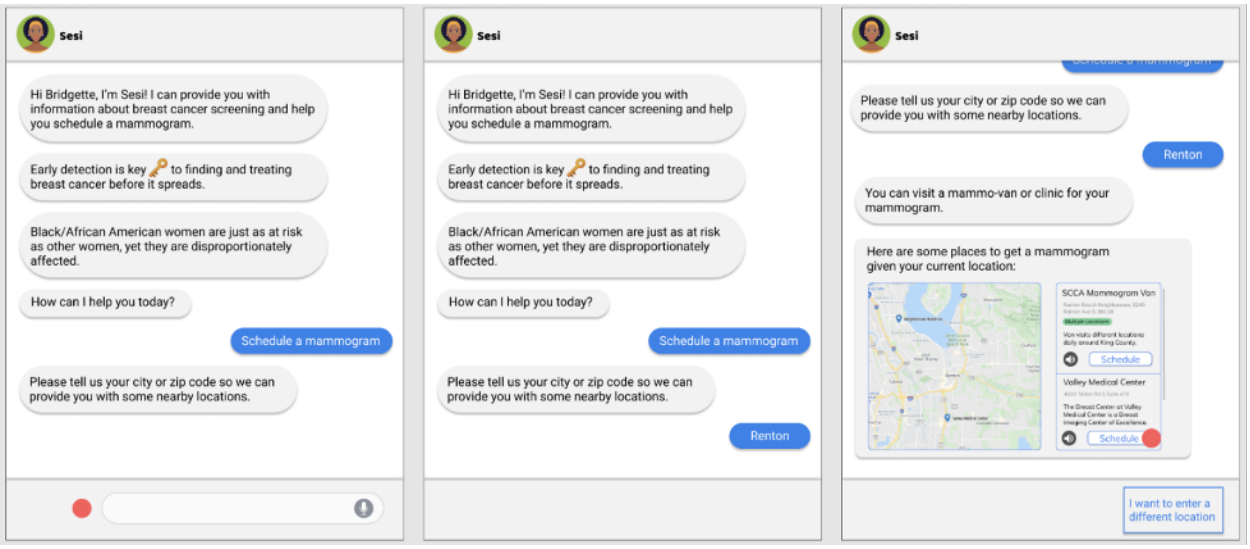


Figure 5.2: Early prototype of the chatbot tool.

with pre-defined domains derived from focus group questions and interview themes to analyze focus group content. Template analysis is a rapid qualitative analysis approach which can be used with focus group data [Fox et al., 2016]. Template domains were agreed upon by investigators and one investigator then reviewed focus groups and conducted content analysis using templates. The completed templates were summarized in a matrix for data visualization and reviewed by all investigators; any disagreements were addressed and resolved.

5.1.2 Results

Identifying key determinants

In the rapid evidence review, 41 relevant studies were identified out of 114 search results. A narrative synthesis was written summarizing determinants identified in the literature. Determinants identified were cataloged and prioritized based on relevance to the implementation strategy. For example, one study found perceptions of lower quality of care if mammograms were done in a mobile clinic setting; we did not include this as a priority determinant because this would not be particularly modifiable in the chatbot design [Adegboyega et al., 2019]. Priority determinants included barriers such as medical mistrust and facilitators such as having personal or family history of breast cancer and recommendations from primary care providers. One priority barrier that emerged from the rapid evidence review was lack of knowledge about breast cancer screening.

Prior work recommended patient education to explain and help individuals learn about the process of getting a mammogram [Ferreira et al., 2021; Adegboyega et al., 2019; Huq et al., 2022]. This informed our design of the initial mockup and early chatbot prototype as a patient education and scheduling tool.

In the qualitative analysis of interviews and focus groups, we similarly elicited themes regarding determinants. Most of the determinants that emerged from interviews and focus groups were also identified in the evidence review (Table 5.1). Overlapping barriers included lack of resources (e.g., cost, insurance, transportation), anxiety about what to expect, fear about negative outcomes associated with the procedure (e.g., pain), medical mistrust, prior negative experiences with the health system (e.g., experiences of racism), lack of information about breast cancer screening, inadequate preparation (while focus groups framed adequate preparation as a facilitator), lack of discussion with family and friends, and lack of clear recommendation from primary care provider. Facilitators that appeared in the rapid evidence review, interviews and/or focus groups included advocacy from a primary care provider, health-related social support, and family or personal history of breast cancer. Some determinants that arose from the interview and focus group data were not present in the rapid evidence review, but were prioritized given relevance to the implementation strategy. For example, participants identified the time spent to make an appointment and the time until the appointment as moderators to scheduling a mammogram. In terms of initial reactions to the chatbot mock-up, 18 out of 20 participants asked thought that the chatbot would be useful for scheduling (one participant was not asked this question).

In the template analysis of focus groups, we analyzed participant feedback to understand facilitators and barriers to the chatbot implementation strategy (Table 5.2).

Participants expressed that the chatbot provided useful information about breast cancer screening through text content and videos. They appreciated information about the cost of screening and how to access screening. In response to the chatbot messages about rescheduling an appointment, some participants thought the question asking "What prevented you from making your appointment?" sounded judgmental, while some appreciated the chatbot was trying to help reschedule without asking too many questions.

> *"I like that...without getting too nosy about why...we're going to try to take care of this issue that you have, and it went straight to trying to figure out to take care of the issue. Which is missing verbally on a regular call versus an application. So I really like that to figure out okay,*

Table 5.1: Determinants to Breast Cancer Screening from Rapid Evidence Review and Analysis of Qualitative Data

Breast Cancer Screening Determinant∗	Qualitative themes	Representative Quotes
• **Lack of resources (cost, insurance, transportation)** • Conflicts with work and/or other competing priorities • *Lack of primary care provider* • **Anxiety about what to expect** • **Fear about pain, exposure to radiation or other negative outcomes associated with procedure** • **Medical mistrust** • **Prior negative experience including experiences of racism** • **Lack of knowledge about breast cancer screening** • **Inadequate preparation/ information given prior to procedure** • **Lack of discussion with friends and family**	Participants discussed barriers including lack of resources, such as finances, transportation, work conflicts, anxiety about what to expect during the mammogram, and prior painful and/or negative experience.	*"I was in West Seattle....a low-income area. And ... there need to be more resources ... that help out women of color... and explain what mammograms consist of. Talk about the cost of it. Talk about resources that individuals can tap into... to be able to get a mammogram." - 56 year old woman (participant 122053)*
• Tailored information about breast cancer • **Family or personal history of breast cancer (or other cancers) as facilitator to screening**	Participants emphasized the importance of outreach to get information about breast cancer screening to the community. It was mentioned that a barrier to screening is not being aware that it was something they should do.	*"How do you know if you're carrying something around, you're sick and you're not knowing what it is, and when you get to the hospital, they diagnosed ... you. But there's things that you could have done prior, if you was told. Some people don't know how to reach out." - 56 year old woman (participant 122053)*
• **Recommendations from PCP** • **Advocacy (or lack of) from PCP** • Time spent to make an appointment • Time until appointment • **Health-related social support**	Participants discussed lack of or equivocal recommendations by physicians for breast cancer screening, even if they initiated discussion. They also discussed reminders, the time it takes to make an appointment and time until appointment as determinants to scheduling screening.	*"I don't think I have been screened this year because of the COVID-19. I've probably seen an email, which is kind of not really personable... because I think in the past I would've got a call... so it didn't make it as urgent or important at the top of the list." – 52 year old woman (participant 122054)*

∗Plain text determinants emerged from interviews and/or focus group only, *italicized* determinants arose from rapid evidence review only, **bolded** determinants were present in both evidence review and interviews and/or focus groups.

Table 5.2: Focus Group Themes

Domain	Theme(s)	Representative quotes
Motivation to screen	Need to overcome competing priorities Should provide more information about importance of screening in the first encounter so that women are motivated to stay proactive in their healthcare	*"I'm trying to feed my baby. I'm trying to get my kids clothes." – (Participant, FG 2)* *"When people are looking, and older women, we don't have a lot of time to be on the internet a lot of times. And so, we get kind of... we got too many other things to do, so we want the information right now." – (Participant, FG 3)*
Reactions to media	Positive reactions to videos Familiarity – liked seeing people who looked like them in image of a woman getting a mammogram	*"It's like, "Oh, that looks like me. Oh, that looks like somebody I can relate to." – (Participant, FG 1)* *"It's going to be important that whoever is involved in this not only looks the same skin color and ethnicity but age-wise, too, so that makes them more relatable, like someone who has actually had a mammogram themselves or who is old enough that needs one, I think would be important too." – (Participant, FG 2)*
Reactions to content	Appreciated discussion of cost Some thought that question asking what prevented you from making your appointment sounded judgmental, while some appreciated it was trying to help reschedule without asking too many questions	*"No, I couldn't tell you what prevented me from attending the appointment. I just don't like it. You can ask me how can I help reschedule your appointment, or can I help you reschedule your appointment." – (Participant, FG 2)*
Perception of chatbot	Some skepticism in multiple groups regarding chatbot persona	*"My first impression would be does she really know what she's talking about? Because just from the picture, I don't know. Yeah, that's what I think." – (Participant, FG3)* *"We don't just want her to just be a random name on the paper. She needs to represent what she's trying to teach us." – (Participant, FG2)*
Comfortability	Majority of participants thought they would feel comfortable using chatbot although some skepticism about using artificial intelligence	*"Got a problem with that whole Big Brother thing." – (Participant, FG3)*

Table 5.3: Focus Group Themes (cont.)

Domain	Theme(s)	Representative quotes
Trust	Expressed privacy concerns about chatbot Concerns about chatbot expertise	*"I've just got a problem with 'Based on our records...' What you doing, following me?...You don't have no privacy...I just would rather for you to say, 'Hi, I notice you are due for a mammogram.'" – (Participant, FG2)* *"Sometimes when those machines are talking, they only have so much information. And so, I may have questions that Sesi may not be able to answer." – (Participant, FG3)*
Usefulness	Saw chatbot as particularly useful for younger women/ first time screening	*"I'd just like to see more of our daughters and our daughter's friends, just to come together as a group and just have the knowledge, just so we get to tap in on that. You know, we know that they may not know, and their friends may not know. So just kind of, give more of an outlook on everything for them as well." – (Participant, FG1)*
Relatability	Concern about cultural inclusiveness – felt that it wasn't personalized outside of community partner involvement	*"And it just didn't speak to me as being a Black woman. That's what I'm going to say. But, you know, let's just be honest. Who made the app?" – (Participant, FG2)*
Desired content	More information about self-exams More BC data about Black women specifically More information about how to prepare for mammogram	*"I would like to have all of it, even the statistics because for me, I would want to go and encourage someone else to get a mammogram. And sometimes, not a lot of statistics, but just knowing among African Americans, that statistics, because a lot of us don't get mammograms because we've heard about the negative things instead of the positive things. So, yeah, I would want to know all of it." – (Participant, FG3)* *"I used to believe that certain diseases were only for white people." – (Participant, FG2)* *"We got to come to the future, and feel comfortable in talking about our health, our breasts, all types of cancer. So some kind of way in there, explain the reason why women of color are disproportionate in this fight for cancer. Knowledge, communication, openness." – (Participant, FG1)*
Desired features/ functions	Appointment reminders Include ways to make BCS social, e.g., "mammogram parties"	*"You could text it or email it to them. I like my little appointment reminders, because I be forgetting stuff. I don't always put stuff on my schedule, on my phone" – (Participant, FG2)* *"There is a need to open up, to increase our bonding as a sisterhood, to feel comfortable to talk about it so [the risk] won't be disproportionate." – (Participant, FG1)*
Usability	Should be efficient – able to schedule quicker than a phone call Did not want to download an app to use	*"And I can go right here and get it all done and be finished in 15, 20 minutes as opposed to being on the phone a half hour... I would definitely use it" – (Participant, FG3)*

what's the next step? How can we help you? What can we do? Now it's waiting for a call from someone if that's the case so that we can get what's needed. Because I missed that when I first was trying to get a mammogram." - Participant, Focus Group 3

Overall, participants appreciated the purpose of the chatbot, but thought that in many ways it fell short.

I mean because that's what the app is for... To kind of make us feel... to draw us in and make us feel taken care of and informed. Educated. - Participant, Focus Group 2

Participants expressed mistrust in the chatbot persona, questioning the chatbot's reliability and describing privacy concerns and intent. They emphasized the importance of cultural inclusivity and familiarity but did not feel like the chatbot prototype achieved these goals.

I do agree with the fact that it needs to be more culturally inclusive and appropriate for us. I didn't feel like it was personalized outside of [B.R.H.'s] involvement, there was nothing that really spoke to our people. - Participant, Focus Group 2

Participants discussed how to improve the chatbot persona and messaging to improve trust and cultural inclusivity. For example, participants suggested providing women with different choices in regards to where to get a mammogram, and not pressuring them to go to a University of Washington facility (e.g., by not associating the chatbot persona with the University of Washington). They also discussed that the chatbot should explain why women of color are disproportionate in the fight against cancer, and acknowledge that women are not always comfortable in talking about their personal health.

The chatbot presented to the focus groups was named "Sesi," which arose from design team meetings with stakeholders. Sesi means "sister" in Sotho, a Bantu language spoken mostly in Southern Africa. Participants expressed frustration about conflating African and Black experience.

Sometimes, because we're Black, other communities patronize on us being Black... they just patronize us as if we know what it is to be in Africa and we don't. We've never been to Africa. We still have the same issues, yes, but we've never been there so we can't relate to certain things or cultures that have because we don't have that. We've never, that was not brought along with us here. - Participant, Focus Group 3

Participants wanted an introduction to Sesi and explanation of the name, for example through the use of an acronym, or preferred the chatbot to have a regular name of a person, such as Ebony. They questioned the value-add of the chatbot presumed to be an app that would require effort to download onto a phone but might only be used once a year. Though participants did think that they would use the chatbot if it could be used to schedule a mammogram more efficiently than by phone.

Mistrust was a main theme present in the rapid evidence review and the qualitative data analyses from the interviews and focus groups. Participants discussed trust and privacy concerns both in the context of interactions with the health care system and the chatbot technology. At the same time, participants noted aspects of the chatbot that increased trust and engagement. For example, they felt reassured to see women who looked like themselves in the chatbot interaction, such as an image of a Black female mammography technician. Given these findings, we decided to focus on optimizing trust in the design of the initial chatbot engagement.

Selecting and Applying Conceptual Frameworks

From our rapid evidence review, qualitative interviews, and focus groups, we identified key determinants – both in the context of breast cancer screening and the chatbot implementation strategy – and hypothesized mechanisms. These insights guided the CPD development and prioritization of trust as a determinant to chatbot use and subsequent breast cancer screening. Using the selected key determinants, we then worked to identify conceptual frameworks based on relevance to and connection of our implementation strategy and proposed mechanism. We used the conceptual frameworks to inform mechanisms through which we hypothesize the implementation strategy to work and moderators which could increase or decrease the mechanism's effect. The CPD which was informed by data from the previous human-centered methods (e.g., interviews, focus groups) which guided our next steps in development of the chatbot prototype – 1) a factorial design experiment measuring trust and engagement with different chatbot personas.

We applied multiple theoretical frameworks regarding trust as a determinant to the evidence-based intervention (breast cancer screening) and the implementation strategy (chatbot) in the development of the CPD. The persuasive health message framework for developing culturally specific messages describes source, channel, and message as distinct components in health messaging and has been used in prior breast cancer

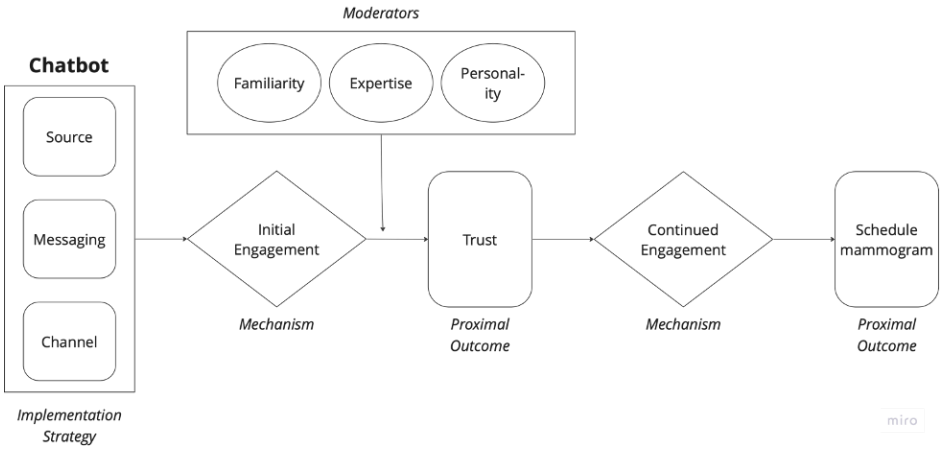


Figure 5.3: Causal Pathway Diagram.

screening campaigns [Hall and Johnson-Turbes, 2015; Witte et al., 1995]. We defined our implementation strategy components using these conventions – source (i.e., chatbot persona – communication style and identity), channel (i.e., form of message delivery, e.g., SMS text), and message (i.e., content of messages). To conceptualize how source (e.g., chatbot persona) may engender trust, we applied a conceptual framework in marketing that identifies expertise, homophily and trustworthiness as characteristics of source credibility [Ismagilova et al., 2020]. Finally, we used a conceptual framework regarding trust in artificial intelligence which includes personality and ability as human characteristics that are important factors that lead to trust in human-robot interactions [Siau and Wang, 2018]. We used the CPD to model how we might address mistrust using initial engagement with the chatbot as a mechanism and trust as a proximal outcome. Using the conceptual frameworks described above, we proposed the moderators to be 1) chatbot expertise, 2) chatbot designed for familiarity (i.e., homophily), and 3) chatbot personality or communication style.

5.2 Optimization Phase

The objective of this phase was to understand how chatbot messages and persona influence trust and engagement. We conducted a randomized factorial experiment to assess the individual components of chatbot persona for breast cancer screening and identify which components have the greatest effect on trust and engagement for Black women. Our identification of key determinants to breast cancer screening from interviews and focus groups were important for identifying trust as an outcome that may lead to screening. These

methods were critical in not only prioritizing trust in the design of the chatbot, but also informing how we might better achieve trust in our design. The Optimization phase uses a multifactorial design to conduct a randomized factorial experiment of specific components identified during the Preparation phase. We then review the results of the factorial experiment to prioritize and discuss adaptations to the chatbot intervention components.

5.2.1 Objectives

Guided by the principles of MOST, we conducted a between-subjects randomized factorial experiment to evaluate design factors hypothesized to promote engagement with a chatbot intervention for breast cancer screening. This study is comprised of an experiment with a 2 (persona: medical provider vs. breast cancer survivor) x 2 (message framing: direct vs. indirect) x 1 (control arm) factorial design. Based on results from human-centered design methods in our existing work, we identified two components of our implementation strategy – chatbot source and messaging – to include in a factorial design experiment to optimize for trust and engagement. We scoped our focus to trust and engagement during the initial exchanges with the chatbot, as studies and our qualitative analysis found that short chatbot interactions [Elsholz et al., 2019] and the chatbot's initial messages and perceptions may significantly impact users' subsequent engagement.

5.2.2 Methods

Factorial Experiment Design

We prioritize and test the chatbot persona as one aspect of the health messaging source. Based on the causal pathway diagram, and congruent with our qualitative analysis and prior literature, both homophily and expertise are important components of engagement and trust in virtual agents [Baylor et al., 2003; Gardiner et al., 2013; Mendu et al., 2018]. In Kim et al., participants suggested a healthcare professional or mother figure as chatbot personas that could engender trust and comfort [Kim et al., 2022]. In our factorial design, we test these characteristics by designing personas with homophily and different types of expertise. The first is a breast cancer survivor – a persona who would have homophily with end-users and expertise through personal experience. The second is a primary care doctor – a persona with expertise as a health care worker as well as homophily with end-users.

We also test different communication styles for the chatbot messaging. Prior work has indicated that natural conversational approaches can be more engaging by using social dialogue. The use of relational behaviors (such as empathy and social dialogue) in a conversational system was shown to increase the desire for continued use [Bickmore et al., 2005]. However, source characteristics of health messages, such as credibility and trustworthiness, may have different levels of influence depending on who is delivering the health information. Black Americans with chronic conditions reported preferring direct communication style in chatbots to appear more factual and less biased [Kim et al., 2022]. In our factorial design, we test direct and indirect chatbot messages drawing from bald-on-record and politeness strategies [Brown et al., 1987]. In the first condition, directness includes commands, direct addresses ("you", 'your"), and maximum restrictions of freedom ("now"). In the second condition, indirectness is characterized by subjunctive modal verb forms ("would like to"), cooperative addresses ("we", "us"), and indirect questions about the learning subject ("would you like..?"). To evaluate whether participants perceive the difference in communication style, we will include semantic differential scales measuring level of directness and indirectness [Lim, 2021; Hu et al., 2022]. Based on prior literature, our hypotheses were as follows:

H1. The representation of the chatbot persona as a Black/African American woman will increase trust and engagement (compared to baseline).

H2. The representation of the chatbot persona as a Black/African American woman will increase intention to use the chatbot in the future (compared to baseline).

H3. The primary care doctor chatbot persona will increase trust and engagement (compared to the breast cancer survivor persona and baseline).

H4. The primary care doctor chatbot persona will increase intention to use the chatbot in the future (compared to the breast cancer survivor persona and baseline).

H5. The direct communication style will increase trust and engagement (compared to indirect communication style and baseline).

H6. The direct communication style will increase intention to use the chatbot in the future (compared to indirect communication style and baseline).

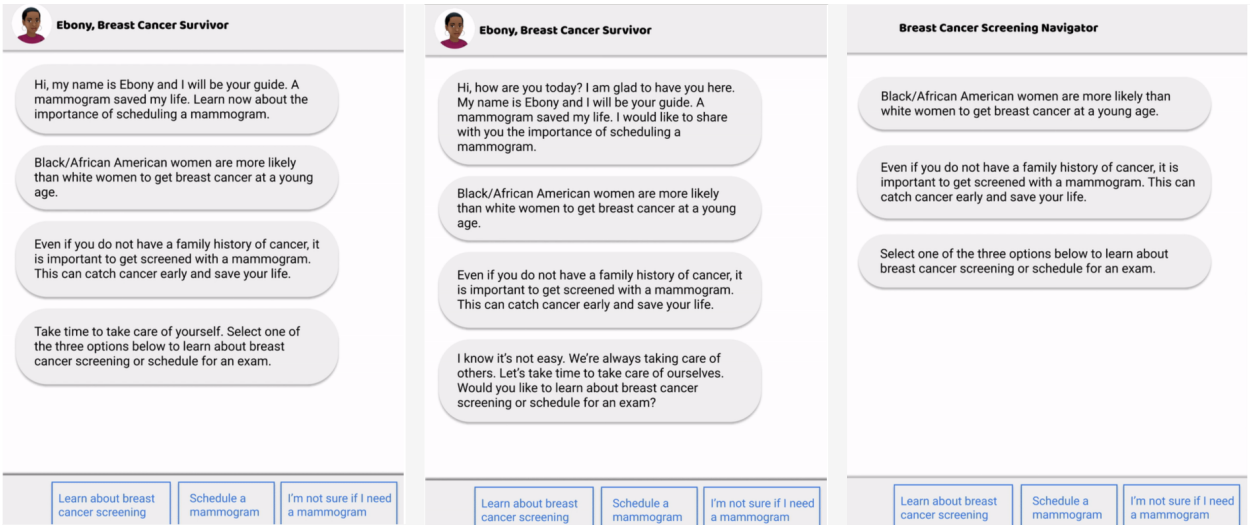


Figure 5.4: Prototype of the chatbot tool showing the breast cancer survivor persona with (1) direct and (2) indirect messaging, and (3) the control condition with no persona or messaging style.

We included two key messages ('Know your risk' and 'Get screened') advised by the Susan G. Komen Breast Cancer Education Toolkit [Komen, 2022] and adapted the chatbot messages in design meetings with team members (Fig. 5.4). The chatbot message delivery is systematically varied across two components, each of which is represented by a separate factor in the 2x2x1 factorial study design with a control arm. Specifically, each participant was randomly assigned to one of five separate experimental conditions. Conditions include: (1) chatbot with a primary care doctor persona and direct communication style; (2) chatbot with a breast cancer survivor persona and direct communication style; (3) chatbot with a primary care doctor persona and indirect communication style; and (4) chatbot with a breast cancer survivor persona and indirect communication style. All participants viewed one condition, and then completed a survey regarding their perceptions about the initial outreach messages from the chatbot and provided demographic information (e.g., age and location). This study was approved by the University of Washington Institutional Review Board.

Outcome Measures

The outcome measures for this experiment were trust, engagement, and intention to use the chatbot for mammography screening. The human-computer trust scale assesses user trust, which is based on similar constructs of trust (benevolence, competence, reciprocity, perceived risk) used in existing trust scales [Jian

et al., 2000]. It is an empirically validated assessment of user trust [Gulati et al., 2019], and uses a 5-point Likert scale from 'Strongly disagree' to 'Strongly agree'. We selected 7 of the 12 items from the human-computer trust scale to create a composite Trust score. The composite Engagement score consists of 4 semantic differential scales assessing traits (important, interesting, relevant, warm) on a 7-point scale [Hollebeek et al., 2014; Aragonés et al., 2015]. We also created a single measure (Intention to Use) to assess participants' likelihood to engage with the chatbot to schedule a mammogram in the future, which is scored on a 5-point Likert scale from 'Very unlikely' to 'Very likely'.

Secondary outcome measures were also included to assess directness and indirectness, expertise, and homophily. To determine how the direct vs. indirect messaging was perceived, we used 7 semantic differential scales assessing traits (direct, friendly, caring, straightforward, demanding, respectful, polite) on a 7-point scale. We included 4 items (Expertise and Homophily) to measure the perceived expertise and similarity of the chatbot on a 5-point Likert scale from 'Strongly disagree' to 'Strongly agree'. Homophily is drawn from a 16-item scale, which consists of four homophily dimensions (attitude, background, value, and appearance) [McCroskey et al., 1975]. We selected 2 items related to the "attitude" dimension. Expertise is drawn from a 4-item scale that measures expertise, experience, knowledge, and qualification [Ohanian, 1990]. For Expertise, we selected 2 items related to expertise and knowledge.

Conventional Content Analysis

A conventional content analysis approach was used to analyze the open-ended survey responses [Hsieh and Shannon, 2005]. The conventional content analysis approach starts with reading all data repeatedly to achieve immersion, and then reading data to derive codes by highlighting text that appears to capture key thoughts or concepts. In the next steps of the process, labels for codes emerge that are reflective of more than one key thought. These codes often come directly from the text and are sorted into categories to develop a codebook scheme. These emergent categories are used to organize and group codes into meaningful clusters. Two members of the research team created an initial codebook using an inductive approach. To create the initial codebook, the two individuals independently coded the open-ended survey responses from a subset of participants related to participants' perceptions of the chatbot. Regular meetings were held among coders to confirm the definition and meaning of new categories and codes. Categories and codes provided

a distinct understanding of the phenomenon, or served to further contextualize and enrich understanding of barriers and facilitators to the chatbot's acceptability. Once the final codebook was created, the survey responses were re-analyzed with the revised codebook. Coders met regularly to ensure consistent application of codes and reconcile differences through discussion or by revisiting transcripts for additional context. Four members of the research team were included in discussions to resolve coding differences in order to finalize the application of codes across transcripts. All coding and analysis were conducted using Microsoft Excel version 16.43.

5.2.3 Results

Participant characteristics

We recruited 550 participants who identified as Black or African American women between the ages of 40-74 years old and were residing in the United States. We deployed the survey on Alchemer and recruited participants on Prolific, an online participant pooling platform. Prolific was used given the platform's ability for selecting the participant population. However, due to the limited number of individuals within the inclusion criteria on Prolific, we also recruited participants from Alchemer Survey Audiences. This platform can be used to recruit pre-screened individuals in the United States based on age, gender, ethnicity and other factors. The survey was deployed over the span of four months (August 22 to December 27, 2022). Some participants were not included in the analysis due to completing the survey in less than half the normal median time (n=28), missing demographic information or not being within the age criteria (n=6), or straightlining survey responses (n=20). Overall, 496 participants were included in the analysis, 327 participants and 169 participants from Prolific and Alchemer, respectively. Participants were compensated $2.50 on Prolific for completing the survey to ensure $15.00/hour compensation, and $3.50 on Alchemer which was determined by the platform based on the survey length.

In the factorial design experiment, participants were on average 52.4 years old (SD=9.1), with the majority residing in the South (62.1% of participants). The participants recruited on Prolific leaned younger than participants recruited on Alchemer, with an average of 50.6 years old (SD=8.1) compared to 56.0 years old (SD=9.8). The survey took participants 10.7 minutes on average to complete (median=6.9).

Table 5.4: Linear regression analysis modeling intention to use the chatbot interface for mammogram scheduling.

	Intention to Use		
	Coef.	SE	p
(Intercept)	3.50	0.11	<2e-16***
Condition (*Peer-Direct*)	-0.14	0.16	0.396
Condition (*Peer-Indirect*)	0.07	0.16	0.643
Condition (*Doctor-Direct*)	0.09	0.16	0.557
Condition (*Doctor-Indirect*)	0.34	0.16	0.037*

Significance: * p<0.05, ** p<0.01, *** p<0.001

Quantitative Findings

The aim of this factorial experiment was to test the effectiveness of chatbot design factors in increasing trust and engagement among Black/African American women. In our data analysis, we used a convergent mixed-methods approach [Fetters et al., 2013; Hong et al., 2017], where both quantitative and qualitative data are collected and then interpreted together. We used this approach to interpret and provide more insight to quantitative findings from the factorial experiment. First, we conducted a linear regression analysis to understand the effectiveness of the chatbot persona and messaging on the primary outcome variables. The primary outcome measures were intention to use the chatbot for mammogram scheduling (Intention to Use), engagement with the chatbot (Engagement), and trust of the chatbot (Trust).

Tables 5.4, 5.5 and 5.12 reflect the study findings related to Intention to Use under our hypotheses **H2**, **H4**, and **H6**. In this study, we hypothesized that the representation of the chatbot persona as a Black/African American woman would increase Intention to Use, compared to the control group (**H2**). To test **H2**, in Table 5.4, we evaluate the relationship between the factorial conditions and Intention to Use, compared to the control group. As seen in Table 5.4, the Doctor-Indirect condition is significantly associated with Intention to Use, with participants in this group having a 0.34 increase in their intention to use the chatbot compared to the control condition (β=0.34, p<0.05). **H2** was partly supported, as one of the conditions (Doctor-Indirect) positively impacted Intention to Use, compared to the control group.

To examine the main and interaction effects of chatbot persona and messaging, we evaluate a full linear regression model, accounting for additional variables, in Tables 5.5 and 5.12. The weights given by the coefficients of a linear regression model measure the effects of the design factors (e.g., persona and mes-

Table 5.5: Linear regression analysis modeling intention to use the chatbot interface for mammogram scheduling.

	Intention to Use		
	Coef.	SE	p
(Intercept)	-0.43	0.45	0.344
Persona (*Doctor*)	0.19	0.09	0.042*
Messaging (*Indirect*)	0.25	0.09	0.007**
Age	-0.01	0.01	0.046*
Comfort with Chatbots	0.45	0.06	2.4e-13***
Comfort with Breast Cancer	0.10	0.05	0.067
Engagement	0.14	0.05	0.007**
Trust	0.34	0.13	0.009**
Homophily	0.06	0.07	0.421
Expertise	0.03	0.08	0.745

Significance: * p<0.05, ** p<0.01, *** p<0.001

saging). In **H4**, we hypothesized that the Doctor persona will increase Intention to Use, compared to the control group and Peer persona, and this hypothesis was supported. In Table 5.5, we found that both the Doctor persona and Indirect messaging are independently and simultaneously correlated with Intention to Use. In this model, the Doctor persona significantly predicted Intention to Use (β=0.19, p<0.05) and Indirect messaging significantly predicted Intention to Use as well (β=0.25, p<0.01). As seen in Table 5.12, we did not observe any interaction effects for the chatbot persona and messaging, thus the interaction variable was removed in Table 5.5. Additionally, we noted that the Doctor persona (in the Doctor-Indirect condition) was associated with intention to use compared to the control group (as seen in Table 5.4). However, in **H6**, it was hypothesized that Direct messaging would increase Intention to Use, compared to the control group and Indirect messaging. This hypothesis was not supported and disconfirmed. Our findings indicated that Indirect messaging was correlated with Intention to Use, when compared to the control group in the Doctor-Indirect condition in Table 5.4 and compared to Direct messaging in Table 5.5. Overall, these findings indicate that a culturally tailored persona, doctor persona, and indirect messaging may lead to a higher intention to use the chatbot for scheduling in the future.

During the analysis, we also observed that comfort interacting with chatbots (Comfort with Chatbots), engagement with the chatbot (Engagement), and trust of the chatbot (Trust) were significantly associated with Intention to Use (Table 5.5). For example, for each 1 point increase in Comfort with Chatbots, the

average Intention to Use increases by 0.45 ($p<0.001$). This observation aligned with our causal pathway modeling where we proposed that trust and engagement will lead to continued engagement with the chatbot. However, we did not find evidence that the chatbot persona and messaging independently impact trust and engagement. Tables 5.6, 5.7, 5.13, 5.8, 5.9 and 5.14 reflect the study findings related to Trust and Engagement under our hypotheses **H1**, **H3**, and **H5**. In this study, we hypothesized that the representation of the chatbot persona as a Black/African American woman would increase trust and engagement, compared to the control group (**H1**). As seen in Table 5.6, the Peer-Indirect condition is significantly associated with Trust (β=0.12, $p<0.05$), compared to the control group. Thus, we found that **H1** was partly supported, as one of the conditions (Peer-Indirect) positively impacted Trust, compared to the control group. There was no significant association between Engagement and the factorial conditions, when compared to the control group (Table 5.8). In **H3**, we hypothesized that the Doctor persona will increase trust and engagement, compared to the Peer persona and control group. Additionally, in **H5**, we hypothesized that the Direct messaging will increase trust and engagement, compared to the Indirect messaging and control group. However, in Tables 5.7 and 5.13 accounting for additional variables, we did not find the main and interaction effects of chatbot persona and messaging to be significantly associated with Trust. We also did not observe any significant independent impact of chatbot persona and messaging on Engagement when examining the main and interaction effects in Tables 5.9 and 5.14. The limited support for hypotheses **H1**, **H3**, and **H5** indicates that a culturally tailored persona, type of persona, and type of messaging do not impact trust and engagement with the chatbot. Though, one finding did indicate that the culturally tailored Peer-Indirect persona was significantly associated with Trust, but not with Intention to Use. A potential explanation is that while the representation of the chatbot persona as a Black/African American woman (in the Peer-Indirect condition) was perceived as trustworthy, there are additional variables outside of trust which influence intention to use the chatbot.

In addition to the main hypotheses, we noted significant findings for the secondary outcome variables, including Homophily and Expertise, and observed that all of the conditions were significantly correlated with homophily, compared to the control group (see Table 5.10). All of the conditions, except for Peer-Direct, were significantly associated with Expertise (see Table 5.11). The lack of significant main effects for chatbot persona and messaging on Trust and Engagement is worth highlighting. It is possible that chatbot

Table 5.6: Linear regression analysis modeling Trust.

	Trust		
	Coef.	SE	p
(Intercept)	3.41	0.04	<2e-16***
Condition (*Peer-Direct*)	0.05	0.06	0.420
Condition (*Peer-Indirect*)	0.12	0.06	0.046*
Condition (*Doctor-Direct*)	0.06	0.06	0.306
Condition (*Doctor-Indirect*)	0.10	0.06	0.130

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.7: Linear regression analysis modeling Trust.

	Trust		
	Coef.	SE	p
(Intercept)	1.92	0.15	<2e-16***
Persona (*Doctor*)	-0.02	0.04	0.578
Messaging (*Indirect*)	0.05	0.04	0.213
Age	0.01	0.00	0.032*
Comfort with Chatbots	0.01	0.02	0.556
Comfort with Breast Cancer	0.05	0.02	0.039*
Homophily	0.15	0.03	7.9e-9***
Expertise	0.18	0.03	6.0e-9***

Significance: * p<0.05, ** p<0.01, *** p<0.001

persona and messaging do not have a strong impact on trust and engagement in the first impression, and interaction with the chatbot is necessary to develop trust and engagement. These findings are consistent with previous research that has shown that interaction with AI systems, such as virtual agents, is an important antecedent to the development of trust [Glikson and Woolley, 2020]. Trust in this study was operationalized as constructs of trust: benevolence, competence, reciprocity, and perceived risk. One limitation to this work is that participants only viewed the initial chatbot messages and did not have the opportunity for further interaction. Participants may have needed more time and a more interactive modality to evaluate the chatbot's capabilities, reciprocity and perceived risks.

Qualitative Findings

Building on the quantitative findings, our qualitative analysis suggested that design implications for the chatbot may be to prioritize (1) a more authentic chatbot persona, and (2) the use of indirect messaging

Table 5.8: Linear regression analysis modeling Engagement.

	Engagement		
	Coef.	SE	p
(Intercept)	5.77	0.11	<2e-16***
Condition (*Peer-Direct*)	0.08	0.16	0.617
Condition (*Peer-Indirect*)	0.13	0.15	0.379
Condition (*Doctor-Direct*)	0.30	0.15	0.055
Condition (*Doctor-Indirect*)	0.24	0.16	0.120

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.9: Linear regression analysis modeling Engagement.

	Engagement		
	Coef.	SE	p
(Intercept)	1.10	0.45	0.014*
Persona (*Doctor*)	0.15	0.09	0.098
Messaging (*Indirect*)	-0.03	0.09	0.728
Age	0.00	0.01	0.958
Comfort with Chatbots	0.25	0.06	1.7e-5***
Comfort with Breast Cancer	0.13	0.05	0.017*
Trust	0.61	0.13	1.7e-6***
Homophily	0.34	0.07	4.4e-7***
Expertise	0.05	0.08	0.534

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.10: Linear regression analysis modeling Homophily.

	Homophily		
	Coef.	SE	p
(Intercept)	2.68	0.09	<2e-16***
Condition (*Peer-Direct*)	0.42	0.13	8.9e-4***
Condition (*Peer-Indirect*)	0.45	0.12	2.2e-4***
Condition (*Doctor-Direct*)	0.36	0.12	0.004**
Condition (*Doctor-Indirect*)	0.39	0.13	0.002**

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.11: Linear regression analysis modeling Expertise.

	Expertise		
	Coef.	SE	p
(Intercept)	3.30	0.07	<2e-16***
Condition (*Peer-Direct*)	0.26	0.11	0.019*
Condition (*Peer-Indirect*)	0.07	0.11	0.520
Condition (*Doctor-Direct*)	0.23	0.11	0.031*
Condition (*Doctor-Indirect*)	0.28	0.11	0.010**

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.12: Linear regression analysis modeling intention to use the chatbot interface for mammogram scheduling, with interaction effect.

	Intention to Use		
	Coef.	SE	p
(Intercept)	-0.43	0.45	0.348
Persona (*Doctor*)	0.19	0.13	0.155
Messaging (*Indirect*)	0.25	0.13	0.057
Persona (*Doctor*)*Messaging (*Indirect*)	0.00	0.19	0.996
Age	-0.01	0.01	0.046*
Comfort with Chatbots	0.45	0.06	2.6e-13***
Comfort with Breast Cancer	0.10	0.05	0.068
Engagement	0.14	0.05	0.007**
Trust	0.34	0.13	0.009**
Homophily	0.06	0.07	0.421
Expertise	0.03	0.08	0.746

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.13: Linear regression analysis modeling Trust, with interaction effect.

	Trust		
	Coef.	SE	p
(Intercept)	1.89	0.16	<2e-16***
Persona (*Doctor*)	0.02	0.05	0.689
Messaging (*Indirect*)	0.09	0.05	0.097
Persona (*Doctor*)*Messaging (*Indirect*)	-0.08	0.08	0.269
Age	0.01	0.00	0.031*
Comfort with Chatbots	0.01	0.02	0.540
Comfort with Breast Cancer	0.04	0.02	0.042*
Homophily	0.15	0.03	1.1e-8***
Expertise	0.18	0.03	4.1e-9***

Significance: * p<0.05, ** p<0.01, *** p<0.001

Table 5.14: Linear regression analysis modeling Engagement, with interaction effect.

	Engagement		
	Coef.	SE	p
(Intercept)	1.07	0.45	0.017*
Persona (*Doctor*)	0.22	0.13	0.100
Messaging (*Indirect*)	0.03	0.13	0.815
Persona (*Doctor*)*Messaging (*Indirect*)	-0.13	0.19	0.497
Age	0.00	0.01	0.948
Comfort with Chatbots	0.25	0.06	1.6e-5***
Comfort with Breast Cancer	0.13	0.05	0.018*
Trust	0.61	0.13	2.1e-6***
Homophily	0.34	0.07	4.9e-7***
Expertise	0.05	0.08	0.496

Significance: * p<0.05, ** p<0.01, *** p<0.001

and healthcare professional persona. In the regression analysis, we observed that the Peer-Indirect and Doctor-Indirect conditions were correlated with trust and intention to use, respectively. Therefore, some of the conditions with the chatbot persona represented as a Black/African American woman influenced trust and intention to use. We also observed that the Doctor persona and Indirect messaging factors were not associated with trust and engagement, but were significantly associated with intention to use the chatbot. In our qualitative analysis, we found that participant may be more likely to use the chatbot for scheduling due to the perception of the Indirect messaging and Doctor persona as caring, professional and realistic. While the regression analysis did not provide explanatory evidence regarding (1) why the culturally tailored persona may lead to trust or intention to use, or (2) why the Doctor persona and Indirect messaging lead to intention to use, the content analysis of survey responses provided insights into participants' perceptions of the chatbot persona and messaging.

Perceptions of Culturally Tailored Chatbot Persona We found that participants desired a chatbot persona that represented a Black/African American woman. Participants in the Peer-Direct (13%), Peer-Indirect (8%), Doctor-Direct (6%), and Doctor-Indirect (6%) conditions liked the "Homophily" of the chatbot, compared to the Control (0%). We found that participants appreciated seeing themselves represented in the chatbot persona, compared to those in the control group. In the Peer-Direct group, participants mentioned that they liked the chatbot because "she represents me being a African American woman" and "it was

presented nice, the avatar was black with a cultural name, so it was very relatable." In the Peer-Indirect group, one participant added "I liked the way that the chatbot was presented because they made her African American. Finally someone with melanin features." Some participants explained that they would be more comfortable talking with the chatbot because it was African-American. While, in the control group, one participant desired a chatbot persona, saying "I think that the chatbot should have an African American avatar instead of just words. I think that it would be more engaging." Overall, participants discussed liking that the chatbot was a female of color with a cultural name, and the statistics about breast cancer in Black women. Participants also felt seen when the chatbot acknowledged that it can be hard to practice self-care.

> *I also liked that it acknowledged that Black women tend to think they don't have time to do many things for themselves due to our many responsibilities when it said that making time could be hard, but we needed to try to make time for ourselves.* - Participant, Doctor-Indirect

Participants added that they wanted the chatbot to not only discuss statistics, but also the environmental, socioeconomic, and systemic issues that contribute to high breast cancer mortality rates.

> *I like that they used an African American avatar for the chatbot. That makes the message seem more customized to me. I disliked that there wasn't an option to just discuss breast cancer itself. That would be the number one question a woman would have to understand why it occurs so often in African American women.* - Participant, Doctor-Direct

Our findings indicate that an important design implication is to incorporate a more authentic chatbot persona, that is culturally tailored without reinforcing stereotypes. Many participants discussed how the chatbot name was stereotypical, and over-emphasized its cultural identity. One participant discussed how the chatbot name made them feel uneasy and affected the level of trust they had in the experience.

> *I think the chatbot being a Black woman is ok. I was uncomfortable with the name "Dr. Ebony", though. I feel as if she would have been just as effective if she were named "Dr. Melissa" or "Dr. Joan".* - Participant, Doctor-Direct

Personas should be designed with care and feedback from many potential users, as chatbots with characteristics including gender, age, and ethnicity identities may risk reinforcing stereotypes [Marino, 2014].

Some participants disliked that the chatbot was solely directed towards Black women, when breast cancer affects all women. Participants said the chatbot doesn't represent all Black women and made them feel uneasy as if they are being targeted.

> *I appreciate the fact that the chatbot is designed to look as closely to Black people as possible, yet it felt a little stereotypical. All women of color don't have natural hair with big, hooped earrings. I hope that the creator of the bot would take the appearance of ALL women of color into consideration.* - Participant, Doctor-Indirect

While a personal identity for chatbots may establish common ground with participants who like seeing themselves represented in the chatbot persona, the persona reinforced stereotypes through the use of the name and image used. While some participants liked the chatbot image, one participant said the use of a real photo of a Black female healthcare provider, rather than the animated chatbot image, would make the chatbot more personable and professional.

Indirect vs. Direct Messaging Participants who viewed the Indirect messaging described the chatbot as being warm, caring, and friendly. In the content analysis, 15 out of 94 participants (16%) in the Doctor-Indirect condition and 10 out of 104 participants in the Peer-Indirect (10%) condition perceived the chatbot to be "Warm, Caring, Friendly", compared to the Doctor-Direct (8%), Peer-Direct (6%), and Control (2%). In the Doctor-Indirect group, participants said they liked the way the chatbot was presented and described the chatbot as "very relatable, non-judgmental and friendly." Additionally, in the Peer-Indirect group, one participant described the chatbot as "not demanding, but gives you something to think about" and "gives a warm welcoming experience." Participants liked that the tone felt conversational and that the chatbot delivered information without being didactic.

> *The avatar is cute and quite representative. The statements are uncomplicated and presented in an inoffensive and "down to earth" way. It presents as a casual conversation, friendly, simply reminding women of the importance of taking care of themselves.* - Participant, Doctor-Indirect

However, one participant in the Doctor-Indirect group recommended that the chatbot could keep the social dialogue, but does not need to ask the question "how are you today?" since it does not seem interested:

"I know it's rather common, but I don't like being asked how I'm doing when the 'person' doesn't actually want to know. I prefer nothing or a 'I hope you're doing well today' type of greeting." In the control group, participants commented that the chatbot messaging felt impersonal and disliked how the chatbot lacked in conversational skills.

> *I did not like that the chatbot seemed so cold. Not warm at all. It could have started out with "Hi" or "Hello". Something that sounds a little more caring.* - Participant, Control

Additionally, participants who viewed the Direct messaging noted that the chatbot "seemed cold, too direct" and "could have had a friendlier introduction." Overall, these participants described the chatbot as boring, robotic, and cold. As prior work has indicated, when a system is represented by a human-like agent lack of empathy could lead to negative user experience and worsen the user's attitude towards the system [Nguyen and Masthoff, 2009]. Chatbot messaging that is indirect and includes social dialogue may be preferred by participants because it appears friendly and caring. This may be an important design consideration particularly for chatbots that discuss personal health matters (e.g. breast cancer) with individuals.

Doctor vs. Breast Cancer Survivor Peer Persona Participants who viewed the Doctor persona described the chatbot as being professional in its presentation of breast cancer information. 5 out of 94 participants (5%) in the Doctor-Indirect condition and 5 out of 97 participants (5%) in the Doctor-Direct condition perceived the chatbot to be "Professional", compared to the Peer-Indirect (3%), Peer-Direct (0%), and Control (1%). In the Doctor-Direct group, participants perceived the chatbot to be "very professional and had valuable information." While one participant in the Peer-Indirect group said they felt like the Peer persona was a friend, another participant expressed that the persona could be better represented as a healthcare professional. One participant noted that they found the educational information to be helpful, but if the focus is on scheduling an appointment, the information could be moved to a sidebar and the persona would be more effective as a healthcare administrator:

> *I like the idea of the chatbot, but...the chatbot should focus on helping the person set up an appointment and attempt to make that conversation sound/feel as natural as possible. The option of more facts or information can be offered at the bottom or via linked words in the*

conversation...if the main focus is booking an appointment, make the chatbot sound like an admin or a front desk personnel. Overall, I love the idea!! - Participant, Peer-Indirect

The use of a healthcare professional for the chatbot persona might help to make the conversation feel more natural, as if they are really setting up an appointment. We found that the Doctor persona was significantly associated with Intention to Use, and this may be due to participants perceiving the persona as more professional and appropriate for the scheduling task. The Doctor persona was also perceived to be more realistic in its language. Participants found it unrealistic for a chatbot to say "A mammogram saved my life" in the Peer persona, compared to "A mammogram can save your life" in the Doctor persona.

I do not like that the chatbot says having a mammogram saved her life. She is just a chatbot. Instead it should 'my name is Ebony and I had breast cancer, but I am using a chatbot to explain why I think you should listen to this message.' - Participant, Peer-Indirect

Participants suggested the chatbot language shouldn't try to imitate a real person, saying a screening saved its life made the chatbot feel a little untrustworthy. Overall, the Doctor persona may be viewed as more professional and realistic because it aligns with the chatbot scheduling task, without trying to closely imitate a real person in its self-introduction.

5.3 Discussion

In this chapter, we elicited feedback from Black/African American women who qualify for breast cancer screening regarding breast cancer screening determinants and experiences as well as feedback on an initial mock-up of a chatbot. We conducted 21 qualitative interviews with Black women living in King or Pierce counties in Washington to understand their perceptions about chatbot outreach for breast cancer information and mammogram scheduling. Qualitative data analysis resulted in themes that were used to develop an early chatbot prototype. This prototype was discussed in three focus groups with 9 Black women. Participants thought that the mock-up of the chatbot tool would be useful for Black women breast cancer screening outreach, but they discussed the importance of designing the chatbot to promote trust and engagement. We identified engagement and trust as two main optimization objectives through the interviews, focus groups, survey and a rapid evidence review. We also identified chatbot messages and persona as primary component

areas to tailor. We tested these components in a factorial design experiment and found that a primary care doctor persona and indirect messaging led to higher intention to use the chatbot for mammogram scheduling.

Interdisciplinary teams should take care to consider how chatbot characteristics may perpetuate stereotypes, especially taking into consideration the source behind the chatbot. As one participant discussed, while the chatbot name and image may not be problematic by themselves, such as with the proposal to create the "Ebony Alert" system for missing Black children and young women [Franklin, 2023], placed together and coming from a team of researchers to encourage breast cancer screening may appear patronizing, rather than uplifting. The design of health technology interventions should actively involve community partners and stakeholders throughout the design process. The factorial experiment demonstrates the effectiveness of indirect messaging and a medical professional chatbot persona as strategies to increase engagement with a chatbot intervention, resulting in greater overall intention to use the chatbot for mammogram scheduling. However, feedback from stakeholders was critical as the qualitative findings from the open-ended survey responses highlighted *why* these experimental factors may have led to higher intention to use, and how they may need to change. They also revealed design considerations in regards to culturally tailoring the chatbot persona. The use of personas (or having some personality) may be helpful for engaging some users, but care should be taken when designing with homophily in mind, especially when designing with marginalized communities. There was a tension between participants feeling tailored to versus targeted by the chatbot design. To address this tension, one consideration for design teams is to continuously engage community partners and stakeholders. Through this engagement, to can be helpful to ask follow-up questions to participants and stakeholders about their feedback. Participants may not share feedback about negative aspects of the chatbot design, or whether they felt targeted, if they are not asked these questions. Therefore, design and research teams should create space for these questions in their interview protocols, and involve community partners as facilitators in the activities to create a comfortable environment to share feedback. Iterating and sharing feedback with partners and stakeholders is needed to assess the culturally tailored design to ensure that it is assets-based, rather than deficit-based, and embraces community values. Through conversations with stakeholders, we found it was important to emphasize self-care and share culturally relevant information about breast cancer screening risks to create an inclusive breast cancer screening outreach tool. Thus, one design recommendation for culturally tailoring is to emphasize relevant information that

stakeholders view as important to share with the community. Culturally relevant information should share specific information, such as statistics about breast cancer, but should avoid over-emphasizing and targeting particular groups (e.g., by referring to 'minority women' compared to Black/African American women). Since spreading awareness was one main goal of the chatbot outreach tool, focusing on communicating this message may be more important, rather than overly tailoring the chatbot persona, particularly when it is based on a fictional character.

The lack of significant main effects for the persona and messaging factors in the linear regression modeling trust and engagement is worth noting. Originally, in our causal pathway diagram, we had included trust as a determinant which leads to the use of a chatbot to schedule a mammogram. Based on the quantitative analysis, it could be the case that trust is not the only factor that leads to intention to use, which may require changes to the CPD. It is possible that trust may not be a determinant because participants did not have enough time for interaction with the chatbot. Additionally, individuals may be willing to use systems they do not trust, if they believe it will lead to positive health outcomes (e.g., getting screened for breast cancer). Though, if trust is one of multiple determinants to intention to use, there may be other factors that influence future chatbot use, such as perceived fit between the persona and task and perception of human-like characteristics. We may need to consider which measures are important to communities when designing CUI-based health interventions. While lack of trustworthiness may not prevent individuals from using the chatbot tool, if the chatbot is perceived as not empathetic or culturally insensitive, this may inhibit long-term use and effectiveness.

Our work in the breast cancer screening outreach context demonstrated that human-centered and implementation science methods - interviews, focus groups, rapid evidence review, causal pathway diagram and factorial design experiment - can work together to identify and understand key determinants to breast cancer screening, and design implementation strategies using conversational user interfaces. To improve engagement in breast cancer screening outreach, we designed the chatbot with a multidisciplinary team and identified key determinants and implementation strategies for the chatbot intervention.

5.4 Summary of Contributions to Thesis

Our work highlights the value of integrating human-centered design and implementation science methods in the development of innovative implementation strategies. We make empirical contributions through studying how conversational user interface design components (e.g., persona and messaging) relate to engagement. By using HCD methods, we were able to understand determinants to breast cancer screening among Black women, and gained important insights about the chatbot as an implementation strategy. Based on the exploratory phase analysis, we identified determinants that could be addressed by a chatbot messaging intervention, e.g. lack of knowledge about breast cancer screening and time spent to make an appointment. While we designed the chatbot with an interdisciplinary team to share culturally relevant information about breast cancer and facilitate scheduling, we found that the early prototype lacked cultural inclusivity and trustworthiness. These insights were integral to the causal pathway diagram development and prioritization of trust as a determinant to chatbot use and subsequent breast cancer screening. The use of chatbot persona and scenarios allowed the team to translate participant feedback into design features, which might not have been possible through group discussion without these materials. Using HCD methods allowed us to bring community partners into early stages of the design process, resulting in concrete feedback that could be incorporated into the next phases. We shared prototypes with stakeholders and participants to iterate on the chatbot design and learn specific feedback on the chatbot components to test in the optimization phase.

In this chapter, we illustrated how human-centered design and implementation science methods were used to 1) identify and prioritize key determinants, 2) select and apply conceptual frameworks, and 3) understand (and design for) strategy mechanisms. We conducted research using human-centered design and implementation science methods to inform and build a CPD to design an implementation strategy. The CPD, which was constructed by data from HCD methods, directly informed our next steps in development of the chatbot prototype – 1) a factorial design experiment measuring trust and engagement and 2) in-progress co-design sessions to iterate on the chatbot messaging. This work adds to existing literature on methods to tailor implementation strategies [Powell et al., 2017; Lewis et al., 2018, 2021; Haines et al., 2021] by proposing an exploratory phase to identify and prioritize determinants, and an optimization phase to guide strategy mechanisms. We propose that an integrated approach of HCD and implementation science methods may help to develop chatbot implementation strategies.

Chapter 6

Discussion and Conclusion

In this chapter, I summarize my thesis contributions, and present discussion points related to the design and research of conversational user interfaces for health interventions.

6.1 Summary of Thesis Contributions

My dissertation research aimed to improve engagement in conversational user interface technologies for health and well-being by integrating human-centered design and implementation science to understand stakeholder needs and optimize intervention components. By drawing from methods in these two disciplines, I discussed how we can improve engagement and better understand how to adapt conversational user interfaces in health interventions.

My dissertation demonstrated this contribution in three research activities: (1) design and evaluation of conversational user interfaces, (2) design and deployment study of chatbot for social needs screening, and (3) design of chatbot for breast cancer screening outreach. I showed that designing new usability heuristics for conversational agents can better support the identification of more usability problems in conversational user interfaces. I proposed 11 heuristics that can be generalized to text, voice and multi-modal conversational agents (Chapter 3). To understand the successes and challenges of implementing conversational user interfaces in health contexts, I applied human-centered design and implementation science methods. I deployed a chatbot and evaluated patients' perceptions of feasibility, acceptability and appropriateness of using a chatbot for social needs screening in the ED. Towards demonstrating that chatbots are an acceptable,

feasible and appropriate form of screening, I leveraged implementation outcome measures and an interview study to reveal individual, contextual, and intervention-related factors that impacted the intervention fidelity. This work drew from the CFIR framework to understand the qualitative data in addition to the outcome measures. The chatbot technology was perceived as responsive, easy to use, efficient, comfortable and enhanced privacy during information disclosure. The qualitative and quantitative data collected from the deployment of this system can be harnessed to validate the design and inform future adaptations of chatbot interventions for social needs screening (Chapter 4). In the context of breast cancer screening outreach, I used human-centered design and implementation science methods to design a chatbot to address health disparities. In the exploratory phase, I evaluated and synthesized stakeholder barriers and facilitators to adoption of the intervention through semi-structured interviews, focus groups, and rapid evidence review. Based on the development of a causal pathway diagram, I then optimized the intervention components through a randomized factorial experiment to understand how a chatbot implementation strategy may lead to future engagement (Chapter 5).

In my work, I developed usability heuristics for conversational user interfaces to characterize and address the challenges that arise in the design of conversational systems. Using two independent health contexts, I demonstrated that the use of human-centered design and implementation science methods can help identify stakeholder needs and challenges that may hinder engagement in health screening and outreach. I identified modifications to chatbot interventions that can be enhanced through the use of implementation science methods to select implementation strategies and improve engagement.

The aforementioned research activities related to conversational user interface design and implementation across two health contexts, therefore, contributed to my thesis:

> *Engagement with conversational user interfaces for health can be improved by integrating human-centered and implementation science methods to understand stakeholder needs and optimize implementation strategies.*

In summary, my dissertation demonstrated that human-centered design and implementation science methods can be used to reveal stakeholder needs, while also prioritizing intervention components for the specific health activity and context. In my work, an understanding of both the conversational user interface intervention and context has led to the appropriate design and adaptation of conversational user interface

interventions necessary to improve engagement in health contexts.

6.2 Design Recommendations and Future Directions

My work across two health contexts provided design recommendations and future research directions in each of their respective chapters. To conclude the thesis, I reflect on my dissertation work and present discussion points that may be useful for the design and research of conversational user interface health interventions.

6.2.1 Design for innovation, and aim for implementation fidelity

Through my work, I have found that human-centered design and implementation science can work together to provide complementary methods that involve stakeholders (e.g., patient, provider, clinic, organization, community) as full partners in the research process from study design through analysis. Human-centered design can be particularly useful in the early stages of research due to its focus on ideating innovative solutions in partnership with stakeholders. In designing effective interventions, it is important to focus on the technology solution and its usability. Heuristic evaluation is a common technique for evaluating user interfaces, but it may need to be adapted for different technologies to identify more usability issues specific to those technologies. The 10 original Nielsen's heuristics for graphical user interfaces (GUIs) may not be appropriate for all kinds of interfaces. For conversational interfaces, we found that some of the original heuristics may be less meaningful (e.g., what does it mean to have searchable help documentation for conversational interfaces?), and new heuristics may be needed to account for the conversational interactions. In Chapter 3, I propose the inclusion of new heuristics for conversational interfaces to address issues of context preservation, trustworthiness, visibility of system status, help and guidance, and error handling. Conversational agent heuristics may support users' implicit expectations and ensure the conversational system does not mislead users about its identity, nor withhold important information about how user data will be used (e.g., Context Preservation and Trustworthiness). These new heuristics can be evaluated for fully functioning chatbots or even low-fidelity prototypes.

In the design process of CUIs it is recommended to begin with defining the purpose of the CUI and prototyping the initial interaction. In Chapter 3, I observed that the first interaction may impact users' expectations about the CUI's capabilities and what they can achieve through the interaction. Starting with

the design of the CUI's introduction, personality, and overall functionality can inform the next steps of designing the dialogue flow. Through this work, I found as well that usability is an important consideration for the design of CUIS, and that new heuristics can highlight usability problems. Yet, I also found that usability is not the only consideration, and that issues related to trustworthiness are also important, even if they are rated as less severe usability issues. Further, attention to contextual adaptation is also necessary for intervention uptake of CUIs. Implementation science approaches may help to guide the design and development of CUIs, particularly in the later stages when an implementation strategy is selected and design components are prioritized. For example, while the conversational agent heuristics were used to inform the design of the social needs screening chatbot, we observed implementation challenges in Chapter 4 related to individual preferences and comfort using the chatbot in the ED context. Future research in human-centered design may consider and benefit from the use of implementation science methods to guide the design of evidence-based interventions and support longitudinal deployments.

6.2.2 Consider contextual fit of the intervention

When designing conversational user interfaces for health interventions, it is important to consider contextual fit and appropriateness. One challenge to implementation is that patients' trust can vary between contexts. In the context of social needs screening, chatbot interventions can be improved by establishing trust through screening in additional contexts outside the emergency department. In emergency departments, providers have reported discomfort asking SDoH screening questions they believed to be stigmatizing, and patients questioned the purpose of the screening questions [Wallace et al., 2020]. Universal screening in primary care may also be conducive to social needs screening as prior research has shown little provider and patient discomfort with SDoH screening in primary care settings [LaForge et al., 2018; de la Vega et al., 2019] and that open discussions of social needs improved patients' relationships with their healthcare team [Drake et al., 2021]. However, it is important to evaluate the patient population to determine how to reach patients facing social needs before implementing screening interventions. Self-administered screening for social needs in primary care settings is generally associated with high levels of acceptability by patients [Gottlieb et al., 2014; Hassan et al., 2013], but healthcare stakeholders have expressed concern about the presence of few patients with social needs in primary care clinics which serve insured members who may be of higher

socioeconomic status [Sundar, 2018]. I believe that the ED waiting area is an ideal location for social needs screening because idle time is spent there, many patients with social needs are present, and the patients with lower ESI who would be more receptive to participating make up the waiting room population. Further research is needed to understand how to improve patient trust in screening technologies in ED environments. For example, future work may consider what social needs information patients are willing or not willing to share with ED providers, and how patient trust varies between contexts.

A second challenge in implementing health interventions is bringing stakeholders into communication with each other in the design process. For example, while I sought to understand patient perceptions of chatbot use for social needs screening, the chatbot intervention may not address different concerns between doctors and patients. Healthcare providers may prefer to better understand patients' social needs through face-to-face conversation and tracking their hospital visits, which may conflict with the needs of some patients who do not want to disclose social needs information. Additionally, the implementation should consider the role of healthcare providers. The perspectives of healthcare professionals, such as ED providers and social workers, are critical to inform the design and future real world implementation of CUIs in healthcare. Our findings suggested that healthcare providers may benefit from training to build rapport with patients and learn about actionable steps they can take in response to patient disclosure/non-disclosure of social needs. There is a need to build buy-in among providers and prepare them through training, as they take on many roles and face a large number of pilots/interventions, particularly at large public hospitals such as Harborview Medical Center. While our collaborators interviewed ED providers and social workers about the chatbot intervention, this dissertation does not include their work on the perspectives of healthcare professionals which would be integral to an understanding of the chatbot as an implementation strategy. Ultimately, attention to contextual fit and diverse stakeholders can lead to more effective implementation of CUI-based health interventions.

6.2.3 Design for community-led health interventions

In addition to considering the contextual fit of interventions, my work examined the impact of technology on people's lives through deployment in real-world contexts. By grounding technology design in people's everyday experiences, the design may be better tailored to individual contexts compared to design inter-

ventions that are based on universal behavior change theories [Michie et al., 2018]. During my research, I found that individuals may lose trust and disengage from technology when it fails to address their needs. For example, users may have health-related questions that may be out of scope of the system's knowledge, as dialogue content for health CUIs is often scripted. Yet, there may be discrepancies and uncertainty in clinical guidelines [Woolf et al., 1999], and guidelines from the medical community may not account for health disparities [Aggarwal et al., 2022]. There is a lack of guidance on how to approach these uncertainties in conversational interactions (e.g., when the user asks a question that is outside of or conflicts with the agent's knowledge base). While I aimed to build trust in conversational user interfaces through new heuristics and clear communication of the health intervention's intentions, it's essential to consider community definitions of implementation success to positively impact trust and engagement. For instance, integrating patient data collection with clinical notes information in existing technologies like EHRs may be a goal for health system stakeholders to screen as many patients as possible. But patients may not want to participate in data collection, and view implementation success as support in accessing relevant resources. In my observations, some ED patients did not see the personal benefit of screening and were wary of data collection. In Chapter 5, while participants valued the potential of the chatbot to assist minority women with scheduling mammograms, they also discussed the risk of reinforcing stereotypes and privacy concerns. Overall, I found it can be challenging to motivate patients to use chatbots for social needs screening and breast cancer screening scheduling. In both contexts, participants wanted to be convinced of the value of using the system and see evidence that the data collected was used to support their needs and community. Furthermore, it is important to understand the intervention's impact on healthcare providers' daily workflows and how the time required to promote a health intervention may lead to disengagement or burnout. In recent work, Shelton et al. state that "implementation research promoting health equity requires foundational and ongoing self-reflection, accountability, and attention to racial equity" [Shelton et al., 2021]. They describe the need for researchers to partner with stakeholders to identify existing community-defined evidence when examining promising practices and interventions for addressing the health effects of structural racism. Human-centered design methods that center stakeholders in the design process may be suited to aid implementation science research in embracing community-defined evidence. In future research at the intersection of human-centered design and implementation science, designers and researchers should reflect upon questions such as: What counts

as evidence? How do we select and prioritize interventions? Who is involved in the development and selection of the intervention? Further research is recommended to increase focus on community-defined evidence and recognize local knowledge in defining effectiveness outcomes [Brownson et al., 2022].

6.2.4 Reflection on the integration of human-centered design and implementation science

To conclude, I reflect on my experiences as a human-centered designer learning and drawing from implementation science methods. In Chapter 5, we applied human-centered design processes in the interview and focus group data collection and analysis. We used qualitative research methods to understand stakeholders' needs and actively involve them as participants in the research and chatbot design. Through the interview and focus group guides, we aimed to understand participants' perspectives by asking about their barriers and facilitators to screening, and follow up on how and if a chatbot could help to address screening barriers. The interview and focus group guides incorporated prototypes to generate conversations about the chatbot pain points, effectiveness, and trustworthiness. These conversations with participants led to redesign ideas regarding the chatbot persona and language. Participants brainstormed alternative names for the chatbot, as well as how the language could be rephrased. This provided concrete feedback that we discussed with our design team and built into the next chatbot prototype. As one example, a focus group participant suggested the chatbot should not ask "How can I help you today?" which appeared patronizing. Instead the chatbot should directly show the available options and how it can help users. This type of concrete feedback informed the chatbot design for the factorial design experiment. In Chapter 5, we used human-centered design processes to not only actively involve participants, but involve our design team, including stakeholders from the community and health system. At the conclusion of the interview data collection and analysis, we held an ideation workshop with the research team and Breast Health Equity committee. During this workshop, we used a 2x2 prioritization matrix as a tool to identify the most impactful and feasible ideas for a screening outreach tool. We asked design team members to vote on the most important barriers to address, which arose during the interviews. Design team members then collaborated and arranged the important barriers ranging from 'low impact' to 'high impact' and 'easy' to 'hard'. This ideation process helped us to actively involve multiple stakeholders in deciding the project's design direction, based on participant feedback, to identify a high impact and feasible idea. The design team was also involved in providing feedback and

designing the chatbot persona and messaging for the focus groups. Their involvement and insight helped us to narrow down what barriers we could address and the most impactful way to address those barriers, from a wide array of potential ideas. For example, connecting patients online with providers who are advocates for breast cancer was deemed as impactful, but challenging to implement. Drawing from human-centered design approaches in this early stage allowed us to spend time ideating and understanding participants' experiences and perspectives on a chatbot tool. We were able to discuss the design tradeoffs of different tools within our team, and start prototyping the chatbot before moving into development and testing. The insights from these human-centered design approaches may have been missed if we had only used the rapid evidence review before formalizing the chatbot design components to optimize.

As a human-centered designer, I found it helpful to draw from human-centered approaches in the research and ideation phases, and to incorporate implementation science perspectives as well. In Chapter 5, I used implementation science approaches in the rapid evidence review and factorial design experiment. In Chapter 4, I also incorporated implementation outcome measures in a real-world deployment and the CFIR implementation science framework to guide patient interviews. These methods were useful in helping to move towards an established evidence base for CUI design in these health contexts. Prior work has indicated that there is a lack of evidence for health interventions in HCI [Klasnja et al., 2017]. In Chapter 4, we used validated implementation outcome measures with the goal of contributing to an evidence base for CUIs in health interventions. We also used the CFIR framework to guide conversations with patients to better understand how to improve the contextual fit of the chatbot for social needs screening. In Chapter 5, the use of the factorial design experiment was integral to providing guidance on CUI design for breast cancer screening outreach. While the interviews, focus groups, and design team collaboration provided insights into the chatbot design, we still had a number of design questions and were unsure about how to present the chatbot persona and design culturally appropriate language. The factorial experiment allowed us to involve a large number of participants to identify the most effective persona and messaging. There are a few recent frameworks that align with our approach and have extended the research phases from MOST [Kowatsch et al., 2019; O'Hara et al., 2022]. Our work in Chapter 5 is similar to these frameworks which utilize design practices in the Preparation phase of the MOST framework, though one difference in our work is the continuation of design and iteration beyond the Optimization phase. While we identified effective components

for the chatbot implementation strategy, we observed based on participant feedback that adjustments to the persona and messaging were necessary to improve trust.

In my research, I observed a number of limitations to our approach in integrating human-centered design and implementation science. One limitation may have been the use of multiple implementation outcome measures in Chapter 4. The measures of acceptability, feasibility, and appropriateness are correlated [Weiner et al., 2017], thus a single measure could have been sufficient to reduce the survey length. However, I found that defining these measures and using them as guides for the interviews was helpful in structuring my conversations with patients. Participants agreed that the chatbot was necessary for social needs screening and appropriate in the ED context since many patients present with social needs. By focusing on acceptability in the interviews, we found that participants had varying preferences on whether they liked the chatbot and were able to discuss how to improve acceptability with participants. In Chapter 5, I also found the use of multiple methods in the exploration phase may not have been an effective approach. As we decided on which determinants to breast cancer screening to prioritize in the chatbot outreach, we engaged in a rapid evidence review, interviews, and focus groups. Both the interviews and focus groups provided important feedback on the chatbot design, and the rapid evidence review allowed us to quickly review literature to understand existing barriers and facilitators, and how to choose an implementation strategy. However, the exploration phase could have been consolidated and completed in a shorter period of time. For example, while the rapid evidence review elicited a wide range of determinants and breast cancer screening interventions, it may not have been a critical method as some research evidence did not translate to our specific context. The interviews and focus groups were useful in understanding determinants related to the chatbot intervention (e.g., trust and engagement) among women in Washington state who would potentially use the chatbot. Before conducting a rapid evidence review, it is advised that researchers ask themselves, "Is there any intervention-related aspect for which getting more information from the literature would be helpful?" In this case, then it might make sense to search for that information in prior research.

One point of guidance for conducting an integrated approach is to foster collaborations between implementation scientists and human-centered designers. Before delving into different models and frameworks, it may be helpful to consult with implementation scientists and potentially form partnerships, to gain insights into which models may be applicable for the research project. In our work, we received project mentorship

from the Optimizing Implementation in Cancer Control (OPTICC) team that included experts in implementation science. We presented our work to the OPTICC team who provided guidance on conducting research approaches, such as the rapid evidence review, and feedback on our research direction. For example, they guided us on which determinants to prioritize from our rapid evidence review. Given the number of implementation frameworks, guidance on the types of frameworks and specific methods to use is critical for HCI researchers. It is important to assess which methodologies are applicable and useful for the intervention and setting.

Through collaboration, HCI researchers may learn about methods to select and optimize components, such as through causal pathway models and factorial design experiments. Experimental design can assist HCI researchers to make decisions about which components to include and how to combine them in a complex intervention. In addition to confirmatory studies, experimental design may aid researchers in adapting and optimizing the design of interfaces for health interventions. The integration of methods could lead to a synthesized evidence base for HCI researchers to build upon. As Klasnja et al. observe, experimental design can provide a rigorous way to evaluate the efficacy of individual components, while using the types of studies that HCI researchers are accustomed to running [Klasnja et al., 2017]. When there is a lack of evidence or insight, a large scale study (e.g., factorial design) can help to make design decisions. If the technology design impacts multiple levels of stakeholders, involves complex workflows, or focuses on behavior change, implementation science methods may be more relevant. In this case, an implementation phase may be needed to implement the technology on a large-scale and monitor reach, impact and side effects. Depending on the context, strict adherence to implementation science processes may not be necessary. The redesign of user interfaces with a focus on improved usability, rather than uptake and sustainment of use, may primarily draw from human-centered design methods. Though, at the least, this consultation may provide HCI researchers with an implementation science perspective going into the intervention design.

In summary, this dissertation discusses an integrated approach to human-centered design and implementation science to improve the design of conversational user interfaces in health interventions. I explore the design and implementation of conversational user interfaces in two health contexts: social needs screening in the ED, and breast cancer screening outreach for Black/African American women. In Chapter 3, I develop new heuristics for conversational agents and apply these usability heuristics to design a chatbot for social

needs screening and resource provision, which is evaluated in a real-world context in Chapter 4. The results show that while patients perceive the chatbot as an acceptable modality, they may need guidance from healthcare professionals and expressed a desire for provider follow-up after completing the screening. In Chapter 5, I use human-centered design and implementation science methods to design a chatbot for breast cancer screening outreach, with a focus on understanding barriers and facilitators to the chatbot use and breast cancer screening, and optimizing intervention components through a factorial experiment. This work highlights the importance of contextual adaptation throughout the design process and the complementary nature of human-centered design and implementation science methods. In this dissertation, I demonstrate how human-centered design and implementation science methods may be integrated to draw from the strengths of both disciplines in the design of chatbot implementation strategies for health.

Bibliography

Washington 211. 2022. Washington 2-1-1.

Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M Bewick, and Mowafa Househ. 2020. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of medical Internet research*, 22(7):e16021.

Adebola Adegboyega, Adaeze Aroh, Kaitlin Voigts, and Hatcher Jennifer. 2019. Regular mammography screening among african american (aa) women: Qualitative application of the pen-3 framework. *Journal of Transcultural Nursing*, 30(5):444–452.

Nancy E Adler, M Maria Glymour, and Jonathan Fielding. 2016. Addressing social determinants of health and health inequalities. *Jama*, 316(16):1641–1642.

Rahul Aggarwal, Kirsten Bibbins-Domingo, Robert W Yeh, Yang Song, Nicholas Chiu, Rishi K Wadhera, Changyu Shen, and Dhruv S Kazi. 2022. Diabetes screening by race and ethnicity in the united states: equivalent body mass index and age thresholds. *Annals of Internal Medicine*, 175(6):765–773.

Pooja Agrawal, Tzuan A Chen, Lorna H McNeill, Chiara Acquati, Shahnjayla K Connors, Vijay Nitturi, Angelica S Robinson, Isabel Martinez Leal, and Lorraine R Reitzel. 2021. Factors associated with breast cancer screening adherence among church-going african american women. *International Journal of Environmental Research and Public Health*, 18(16):8494.

Ahmed T Ahmed, Brian T Welch, Waleed Brinjikji, Wigdan H Farah, Tara L Henrichsen, M Hassan Murad, and John M Knudsen. 2017. Racial disparities in screening mammography in the united states: a systematic review and meta-analysis. *Journal of the American College of Radiology*, 14(2):157–165.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Jessica S Ancker, Yolanda Barrón, Maxine L Rockoff, Diane Hauser, Michelle Pichardo, Adam Szerencsy, and Neil Calman. 2011. Use of an electronic patient portal among disadvantaged populations. *Journal of general internal medicine*, 26:1117–1123.

Jessica S Ancker, Elizabeth Mauer, Diane Hauser, and Neil Calman. 2016. Expanding access to high-quality plain-language patient education information through context-specific hyperlinks. In *AMIA Annual Symposium Proceedings*, volume 2016, page 277. American Medical Informatics Association.

Denise L Anthony, Celeste Campos-Castillo, and Paulina S Lim. 2018. Who isn't using patient portals and why? evidence and implications from a national sample of us adults. *Health Affairs*, 37(12):1948–1954.

Juan I Aragonés, Lucía Poggio, Verónica Sevillano, Raquel Pérez-López, and María-Luisa Sánchez-Bernardos. 2015. Measuring warmth and competence at inter-group, interpersonal and individual levels/medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. *Revista de Psicología Social*, 30(3):407–438.

Joan S Ash, Marc Berg, and Enrico Coiera. 2004. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, 11(2):104–112.

John Auerbach and Brian Castrucci. 2019. Meeting individual social needs falls short of addressing social determinants of health. *Policy & Practice*, 77(2):25–28.

Marc Auriacombe, Sarah Moriceau, Fuschia Serre, Cécile Denis, Jean-Arthur Micoulaud-Franchi, Etienne de Sevin, Emilien Bonhomme, Stéphanie Bioulac, Mélina Fatseas, and Pierre Philip. 2018. Develop-

ment and validation of a virtual agent to screen tobacco and alcohol use disorders. *Drug and alcohol dependence*, 193:1–6.

Mark S Bauer, Laura Damschroder, Hildi Hagedorn, Jeffrey Smith, and Amy M Kilbourne. 2015. An introduction to implementation science for the non-specialist. *BMC psychology*, 3(1):1–12.

Amy Baylor, E Shen, and Xiaoxia Huang. 2003. Which pedagogical agent do learners choose? the effects of gender and ethnicity. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 1507–1510. Association for the Advancement of Computing in Education (AACE).

Rinad S Beidas, Alison M Buttenheim, and David S Mandell. 2021. Transforming mental health care delivery through implementation science and behavioral economics. *JAMA psychiatry*, 78(9):941–942.

Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, 6(3):4.

Evelyn Berger-Jenkins, Catherine Monk, Katherine D'Onfro, Majeda Sultana, Lisa Brandt, Jyoti Ankam, Nadiuska Vazquez, and Dodi Meyer. 2019. Screening for both child behavior and social determinants of health in pediatric primary care. *Journal of developmental and behavioral pediatrics: JDBP*, 40(6):415.

Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403.

Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, pages 23–54. Springer.

Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30.

Timothy W Bickmore, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow. 2010. Usability of conversational agents by patients

with inadequate health literacy: evidence from two clinical trials. *Journal of health communication*, 15(S2):197–210.

Heather Bleacher, Corey Lyon, Logan Mims, Kathy Cebuhar, and Anowara Begum. 2019. The feasibility of screening for social determinants of health: seven lessons learned. *Family Practice Management*, 26(5):13–19.

Nienke Bleijenberg, Janneke M de Man-van Ginkel, Jaap CA Trappenburg, Roelof GA Ettema, Carolien G Sino, Noor Heim, Thóra B Hafsteindóttir, David A Richards, and Marieke J Schuurmans. 2018. Increasing value and reducing waste by optimizing the development of complex interventions: Enriching the development phase of the medical research council (mrc) framework. *International journal of nursing studies*, 79:86–93.

Daniel S Blumenthal and Ernest Alema-Mensah. 1997. Effect of a cancer screening intervention conducted by lay health workers among inner-city women. *American journal of preventive medicine*, 13(1):51–57.

Sarabeth Broder-Fingert, Jocelyn Kuhn, Radley Christopher Sheldrick, Andrea Chu, Lisa Fortuna, Megan Jordan, Dana Rubin, and Emily Feinberg. 2019. Using the multiphase optimization strategy (most) framework to test intervention delivery strategies: a study protocol. *Trials*, 20:1–15.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Ross C Brownson, Rachel C Shelton, Elvin H Geng, and Russell E Glasgow. 2022. Revisiting concepts of evidence in implementation science. *Implementation Science*, 17(1):1–25.

Christopher Burton, Aurora Szentagotai Tatar, Brian McKinstry, Colin Matheson, Silviu Matu, Ramona Moldovan, Michele Macnab, Elaine Farrow, Daniel David, Claudia Pagliari, et al. 2016. Pilot randomised controlled trial of help4mood, an embodied virtual agent-based system to support treatment of depression. *Journal of telemedicine and telecare*, 22(6):348–355.

Elena D Butler, Anna U Morgan, and Shreya Kangovi. 2020. Screening for unmet social needs: patient engagement or alienation? *NEJM Catalyst Innovations in Care Delivery*, 1(4).

Piotr Calak. 2013. *Smartphone evaluation heuristics for older adults*. Ph.D. thesis.

Roberta Capp, Gregory J Misky, Richard C Lindrooth, Benjamin Honigman, Heather Logan, Rose Hardy, Dong Q Nguyen, and Jennifer L Wiler. 2017. Coordination program reduced acute care use and increased primary care visits among frequent emergency care users. *Health Affairs*, 36(10):1705–1711.

Justine Cassell and Timothy Bickmore. 2000. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56.

CDC. 2022. Leading causes of death – females – non-hispanic black – united states, 2018.

Disability Benefits Center. Benefits eligibility screening tool (best).

Pew Research Center. 2021. Mobile fact sheet.

Benjamin Chaix, Jean-Emmanuel Bibault, Arthur Pienkowski, Guillaume Delamon, Arthur Guillemassé, Pierre Nectoux, Benoît Brouard, et al. 2019. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR cancer*, 5(1):e12856.

Victoria L Champion, Jeffrey K Springston, Terry W Zollinger, Robert M Saywell Jr, Patrick O Monahan, Qianqian Zhao, and Kathleen M Russell. 2006. Comparison of three interventions to increase mammography screening in low income african american women. *Cancer detection and prevention*, 30(6):535–544.

Claire Chang, Christina Ceci, Megha Uberoi, Marika Waselewski, and Tammy Chang. 2022. Youth perspectives on their medical team's role in screening for and addressing social determinants of health. *Journal of Adolescent Health*, 70(6):928–933.

Real Change. 2021. Emerald city resource guide.

Christina Hunter Chapman, Clyde B Schechter, Christopher J Cadham, Amy Trentham-Dietz, Ronald E Gangnon, Reshma Jagsi, and Jeanne S Mandelblatt. 2021. Identifying equitable screening mammography strategies for black women in the united states using simulation modeling. *Annals of internal medicine*, 174(12):1637–1646.

Karen G Cheng, Francisco Ernesto, and Khai N Truong. 2008. Participant and interviewer attitudes toward handheld computers in the context of hiv/aids programs in sub-saharan africa. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 763–766.

Dana E Chisnell, Janice C Ginny Redish, and AMY Lee. 2006. New heuristics for understanding older adults as web users. *Technical Communication*, 53(1):39–59.

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*.

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*, pages 1–12.

Heather Cole-Lewis and Trace Kershaw. 2010. Text messaging as a tool for behavior change in disease prevention and management. *Epidemiologic reviews*, 32(1):56–69.

Linda M Collins, Susan A Murphy, Vijay N Nair, and Victor J Strecher. 2005. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30(1):65–73.

Valire Carr Copeland, Yoo Jung Kim, and Shaun M Eack. 2018. Effectiveness of interventions for breast cancer screening in african american women: A meta-analysis. *Health services research*, 53:3170–3188.

Judith B Cornelius, Janet S St Lawrence, Jacquelyn C Howard, Deval Shah, Avinash Poka, Delilah McDonald, and Ann C White. 2012. Adolescents' perceptions of a mobile cell phone text messaging-enhanced intervention and development of a mobile cell phone-based hiv prevention intervention. *Journal for specialists in pediatric nursing: JSPN*, 17(1):61.

Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, page 43. ACM.

John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Bobby Daly and Olufunmilayo I Olopade. 2015. A perfect storm: how tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA: a cancer journal for clinicians*, 65(3):221–238.

Laura J Damschroder, David C Aron, Rosalind E Keith, Susan R Kirsh, Jeffery A Alexander, and Julie C Lowery. 2009. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation science*, 4(1):1–15.

Claudia M Davis. 2021. Health beliefs and breast cancer screening practices among african american women in california. *International Quarterly of Community Health Education*, 41(3):259–266.

Maria De Jesus, Shalini Ramachandra, Alexis De Silva, Shirley Liu, Ethan Dubnansky, Kingsley Iyawe, Astrid Jimenez, Laura Logie, and MC Jackson. 2021. A mobile health breast cancer educational and screening intervention tailored for low-income, uninsured latina immigrants. *Women's Health Reports*, 2(1):325–336.

Kerstin Denecke, Sandra Lutz Hochreutener, Annkathrin Pöpel, and Richard May. 2018. Self-anamnesis with a conversational user interface: concept and usability study. *Methods of information in medicine*, 57(05/06):243–252.

Carol E DeSantis, Jiemin Ma, Mia M Gaudet, Lisa A Newman, Kimberly D Miller, Ann Goding Sauer, Ahmedin Jemal, and Rebecca L Siegel. 2019. Breast cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(6):438–451.

Maureen Dobbins. 2017. Rapid review guidebook. *Natl Collab Cent Method Tools*, 13:25.

Alex R Dopp, Kathryn E Parisi, Sean A Munson, and Aaron R Lyon. 2020. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. *Implementation Science Communications*, 1(1):1–13.

Connor Drake, Heather Batchelder, Tyler Lian, Meagan Cannady, Morris Weinberger, Howard Eisenson, Emily Esmaili, Allison Lewinski, Leah L Zullig, Amber Haley, et al. 2021. Implementation of social needs screening in primary care: a qualitative study using the health equity implementation framework. *BMC health services research*, 21:1–16.

Martin P Eccles and Brian S Mittman. 2006. Welcome to implementation science. *Implementation Science*, 1(1).

J William Eley, Holly A Hill, Vivien W Chen, Donald F Austin, Margaret N Wesley, Hyman B Muss, Raymond S Greenberg, Ralph J Coates, Pelayo Correa, Carol K Redmond, et al. 1994. Racial differences in survival from breast cancer: results of the national cancer institute black/white cancer survival study. *Jama*, 272(12):947–954.

Danielle Elmasri and Anthony Maeder. 2016. A conversational agent for an online mental health intervention. In *Brain Informatics and Health: International Conference, BIH 2016, Omaha, NE, USA, October 13-16, 2016 Proceedings*, pages 243–251. Springer.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305.

Xiangmin Fan, Daren Chao, Zhan Zhang, Dakuo Wang, Xiaohua Li, and Feng Tian. 2021. Utilization of self-diagnosis health chatbots in real-world settings: case study. *Journal of medical Internet research*, 23(1):e19928.

Cristiana Sofia Ferreira, Joana Rodrigues, Stefanie Moreira, Filipa Ribeiro, and Adhemar Longatto-Filho. 2021. Breast cancer screening adherence rates and barriers of implementation in ethnic, cultural and religious minorities: A systematic review. *Molecular and Clinical Oncology*, 15(1):1–9.

Michael D Fetters, Leslie A Curry, and John W Creswell. 2013. Achieving integration in mixed methods designs—principles and practices. *Health services research*, 48(6pt2):2134–2156.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy

to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Annie B Fox, Alison B Hamilton, Susan M Frayne, Shannon Wiltsey-Stirman, Bevanne Bean-Mayberry, Diane Carney, Brooke AL Di Leone, Jennifer M Gierisch, Karen M Goldstein, Yasmin Romodan, et al. 2016. Effectiveness of an evidence-based quality improvement approach to cultural competence training: The veterans affairs' caring for women veterans program. *The Journal of continuing education in the health professions*, 36(2):96.

Jonathan Franklin. 2023. A california bill would create an alert system for missing black women and youth.

Caroline Free, Gemma Phillips, Louise Watson, Leandro Galli, Lambert Felix, Phil Edwards, Vikram Patel, and Andy Haines. 2013. The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis. *PLoS medicine*, 10(1):e1001363.

Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, Michiel Rauws, et al. 2018. Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, 5(4):e9782.

Lucian Galescu, James Allen, George Ferguson, Jill Quinn, and Mary Swift. 2009. Speech recognition in a dialog system for patient health monitoring. In *2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop*, pages 302–307. IEEE.

Paula Gardiner, Megan B Hempstead, Lazlo Ring, Timothy Bickmore, Leanne Yinusa-Nyahkoon, Huong Tran, Michael Paasche-Orlow, Karla Damus, and Brian Jack. 2013. Reaching women through health information technology: the gabby preconception care system. *American Journal of Health Promotion*, 27(3_suppl):eS11–eS20.

Wambui G Gathirua-Mwangi, Patrick O Monahan, Timothy Stump, Susan M Rawl, Celette Sugg Skinner, and Victoria L Champion. 2016. Mammography adherence in african-american women: Results of a randomized controlled trial. *Annals of Behavioral Medicine*, 50(1):70–78.

Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. *Stud Health Technol Inform*, 252:51–56.

Angela N Giaquinto, Kimberly D Miller, Katherine Y Tossas, Robert A Winn, Ahmedin Jemal, and Rebecca L Siegel. 2022. Cancer statistics for african american/black people 2022. *CA: a cancer journal for clinicians*, 72(3):202–229.

James Glass. 1999. Challenges for spoken dialogue systems. In *Proceedings of the 1999 IEEE ASRU Workshop*.

Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Anupam Goel, Julie George, and Robert C Burack. 2008. Telephone reminders increase re-screening in a county breast screening program. *Journal of Health Care for the Poor and Underserved*, 19(2):512–521.

Rachel Gold, Arwen Bunce, Stuart Cowburn, Katie Dambrun, Marla Dearing, Mary Middendorf, Ned Mossman, Celine Hollombe, Peter Mahr, Gerardo Melgar, et al. 2018. Adoption of social determinants of health ehr tools by community health centers. *The Annals of Family Medicine*, 16(5):399–407.

Google. n.d. Learn about conversation - conversation design. `https://designguidelines.withgoogle.com/conversation/conversation-design/learn-about-conversation.html`.

James A Gordon. 1999. The hospital emergency department as a social welfare institution. *Annals of emergency medicine*, 33(3):321–325.

Laura Gottlieb, Danielle Hessler, Dayna Long, Anais Amaya, and Nancy Adler. 2014. A randomized trial on screening for social determinants of health: the iscreen study. *Pediatrics*, 134(6):e1611–e1618.

Laura M Gottlieb, Holly Wing, and Nancy E Adler. 2017. A systematic review of interventions on patients' social and economic needs. *American journal of preventive medicine*, 53(5):719–729.

Andrea K Graham, Emily G Lattie, Byron J Powell, Aaron R Lyon, Justin D Smith, Stephen M Schueller, Nicole A Stadnick, C Hendricks Brown, and David C Mohr. 2020. Implementation strategies for digital mental health interventions in health care settings. *American Psychologist*, 75(8):1080.

Trisha Greenhalgh, Joseph Wherton, Chrysanthi Papoutsi, Jennifer Lynch, Gemma Hughes, Susan Hinder, Nick Fahy, Rob Procter, Sara Shaw, et al. 2017. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*, 19(11):e8775.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

David Griol and Zoraida Callejas. 2016. Mobile conversational agents for context-aware care applications. *Cognitive Computation*, 8:336–356.

Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015.

Yuqi Guo, Tyrone C Cheng, and Hee Yun Lee. 2019. Factors associated with adherence to preventive breast cancer screenings among middle-aged african american women. *Social work in public health*, 34(7):646–656.

Emily R Haines, Alex Dopp, Aaron R Lyon, Holly O Witteman, Miriam Bender, Gratianne Vaisson, Danielle Hitch, and Sarah Birken. 2021. Harmonizing evidence-based practice, implementation context, and implementation strategies with user-centered design: a case example in young adult cancer care. *Implementation science communications*, 2(1):45.

Ingrid J Hall and Ashani Johnson-Turbes. 2015. Use of the persuasive health message framework in the development of a community-based mammography promotion campaign. *Cancer Causes & Control*, 26:775–784.

Dale Hardy and Daniel Y Du. 2021. Socioeconomic and racial disparities in cancer stage at diagnosis, tumor size, and clinical outcomes in a large cohort of women with breast cancer, 2007–2016. *Journal of racial and ethnic health disparities*, 8:990–1001.

Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Paul A Harris. 2012. Research electronic data capture (redcap)-planning, collecting and managing data for clinical and translational research. In *BMC bioinformatics*, volume 13, pages 1–1. BioMed Central.

Richard Harte, Liam Glynn, Alejandro Rodríguez-Molinero, Paul MA Baker, Thomas Scharf, Leo R Quinlan, Gearóid ÓLaighin, et al. 2017. A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. *JMIR human factors*, 4(1):e5443.

Kimberly R Hartson, Lindsay J Della, Kristi M King, Sam Liu, Paige N Newquist, and Ryan E Rhodes. 2022. Application of the ideas framework in adapting a web-based physical activity intervention for young adult college students. In *Healthcare*, volume 10, page 700. MDPI.

Areej Hassan, Emily A Blood, Aaron Pikcilingis, Emily G Krull, LaQuita McNickles, Glenn Marmon, Sarah Wylie, Elizabeth R Woods, and Eric W Fleegler. 2013. Youths' health-related social problems: concerns often overlooked during the medical visit. *Journal of Adolescent Health*, 53(2):265–271.

Jemma Hawkins, Kim Madden, Adam Fletcher, Luke Midgley, Aimee Grant, Gemma Cox, Laurence Moore, Rona Campbell, Simon Murphy, Chris Bonell, et al. 2017. Development of a framework for the co-production and prototyping of public health interventions. *BMC public health*, 17(1):1–11.

Samantha Hendren, Paul Winters, Sharon Humiston, Amna Idris, Shirley XL Li, Patricia Ford, Raymond Specht, Stephen Marcus, Michael Mendoza, and Kevin Fiscella. 2014. Randomized, controlled trial of a multimodal intervention to improve cancer screening rates in a safety-net primary care practice. *Journal of general internal medicine*, 29:41–49.

Monique M Hennink, Bonnie N Kaiser, and Vincent C Marconi. 2017. Code saturation versus meaning saturation: how many interviews are enough? *Qualitative health research*, 27(4):591–608.

Setia Hermawati and Glyn Lawson. 2016. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied ergonomics*, 56:34–51.

Linda D Hollebeek, Mark S Glynn, and Roderick J Brodie. 2014. Consumer brand engagement in social media: Conceptualization, scale development and validation. *Journal of interactive marketing*, 28(2):149–165.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 207–214.

Quan Nha Hong, Pierre Pluye, Mathieu Bujold, and Maggy Wassef. 2017. Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic reviews*, 6(1):1–14.

Noura Howell, Audrey Desjardins, and Sarah Fox. 2021. Cracks in the success narrative: Rethinking failure in design research through a retrospective trioethnography. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(6):1–31.

Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.

Clarissa Hsu, Stephanie Cruz, Hilary Placzek, Michelle Chapdelaine, Sara Levin, Fabiola Gutierrez, Sara Standish, Ian Maki, Mary Carl, Miriam Rosa Orantes, et al. 2020. Patient perspectives on addressing social needs in primary care using a screening and resource referral intervention. *Journal of General Internal Medicine*, 35:481–489.

Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. 2022. Polite or direct? conversation design of a smart display for older adults based on politeness theory. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Maisha R Huq, Nathaniel Woodard, Leonore Okwara, and Cheryl L Knott. 2022. Breast cancer educational needs and concerns of african american women below screening age. *Journal of Cancer Education*, 37(6):1677–1683.

Apple Computer Inc. 1987. *Apple Human Interface Guidelines: The Apple Desktop Interface*. Addison Wesley Publishing Company.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Keith Instone. 1997. Site usability heuristics for the web. http://instone.org/heuristics.

Elvira Ismagilova, Emma Slade, Nripendra P Rana, and Yogesh K Dwivedi. 2020. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services*, 53:101736.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 895–906.

Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71.

Maria L Jibaja-Weiss, Robert J Volk, Paul Kingery, Quentin W Smith, and J David Holcomb. 2003. Tailored messages for breast and cervical cancer screening of low-income and minority women using medical records data. *Patient education and counseling*, 50(2):123–132.

Sara Johnson, Patrick Liu, David Campa, Charmaine Dorsey, Dennis Hsieh, Social, and Behavioral Determinants of Health Workgroup. 2019. Los angeles county health agency: Social and behavioral determinants of health screening guide.

Claire EL Jones, Jill Maben, Ruth H Jack, Elizabeth A Davies, Lindsay JL Forbes, Grace Lucas, and Emma Ream. 2014. A systematic review of barriers to early presentation and diagnosis with breast cancer among black women. *BMJ open*, 4(2):e004076.

Takeshi Kamita, Tatsuya Ito, Atsuko Matsumoto, Tsunetsugu Munakata, and Tomoo Inoue. 2019. A chatbot system for mental healthcare based on sat counseling method. *Mobile Information Systems*, 2019.

Maria C Katapodi, Penny F Pierce, and Noreen C Facione. 2010. Distrust, predisposition to use health services and breast cancer screening: results from a multicultural community-based survey. *International Journal of Nursing Studies*, 47(8):975–983.

Sabra L Katz-Wise, Sari L Reisner, Jaclyn M White Hughto, and Stephanie L Budge. 2017. Self-reported

changes in attractions and social determinants of mental health in transgender adults. *Archives of Sexual Behavior*, 46:1425–1439.

Steven Kaufman, Nadia Ali, Victoria DeFiglio, Kelly Craig, and Jeffrey Brenner. 2014. Early efforts to target and enroll high-risk diabetic patients into urban community-based programs. *Health promotion practice*, 15(2_suppl):62S–70S.

Junhan Kim, Jana Muhic, Lionel Peter Robert, and Sun Young Park. 2022. Designing chatbots with black americans with chronic conditions: Overcoming challenges against covid-19. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Junhan Kim, Sun Young Park, and Lionel P Robert. 2020. Bridging the health disparity of african americans through conversational agents. *Digital Government: Research and Practice*, 2(1):1–7.

Sage J Kim, Anne Elizabeth Glassgow, Karriem S Watson, Yamile Molina, and Elizabeth A Calhoun. 2018. Gendered and racialized social expectations, barriers, and delayed breast cancer diagnosis. *Cancer*, 124(22):4350–4357.

Predrag Klasnja, Eric B Hekler, Elizabeth V Korinek, John Harlow, and Sonali R Mishra. 2017. Toward usable evidence: optimizing knowledge accumulation in hci research on health behavior change. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3071–3082.

Naomi Y Ko, Susan Hong, Robert A Winn, and Gregory S Calip. 2020. Association of insurance status and racial disparities with the detection of early-stage breast cancer. *JAMA oncology*, 6(3):385–392.

Ahmet Baki Kocaballi, Emre Sezgin, Leigh Clark, John M Carroll, Yungui Huang, Jina Huh-Yoo, Junhan Kim, Rafal Kocielnik, Yi-Chieh Lee, Lena Mamykina, et al. 2022. Design and evaluation challenges of conversational agents in health care and well-being: Selective review study. *Journal of medical Internet research*, 24(11):e38525.

Rafal Kocielnik. 2021. *Designing Engaging Conversational Interactions for Health & Behavior Change*. University of Washington.

Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. 2019. Harborbot: A chatbot for social needs screening. In *AMIA Annual Symposium Proceedings*, volume 2019, page 552. American Medical Informatics Association.

Rafal Kocielnik, Raina Langevin, James S George, Shota Akenaga, Amelia Wang, Darwin P Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T Hsieh, et al. 2021. Can i talk to you about your social needs? understanding preference for conversational user interface in health. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, pages 1–10.

Susan G. Komen. 2022. Breast cancer education toolkit for use with black and african american communities.

Tobias Kowatsch, Lena Otto, Samira Harperink, Amanda Cotti, and Hannes Schlieter. 2019. A design and evaluation framework for digital health interventions. *IT-Information Technology*, 61(5-6):253–263.

Matthew W Kreuter, Celette Sugg-Skinner, Cheryl L Holt, Eddie M Clark, Debra Haire-Joshu, Qiang Fu, Angela C Booker, Karen Steger-May, and Dawn Bucholtz. 2005. Cultural tailoring for mammography and fruit and vegetable intake among low-income african-american women in urban public health centers. *Preventive medicine*, 41(1):53–62.

Kate LaForge, Rachel Gold, Erika Cottrell, Arwen E Bunce, Michelle Proser, Celine Hollombe, Katie Dambrun, Deborah J Cohen, and Khaya D Clark. 2018. How 6 organizations developed tools and processes for social determinants of health screening in primary care: an overview. *The Journal of ambulatory care management*, 41(1):2.

Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

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):1248–1258.

Jennifer Leeman, Sarah A Birken, Byron J Powell, Catherine Rohweder, and Christopher M Shea. 2017. Beyond "implementation strategies": classifying the full range of strategies used in implementation science and practice. *Implementation Science*, 12:1–9.

Cara C Lewis, Peggy A Hannon, Predrag Klasnja, Laura-Mae Baldwin, Rene Hawkes, Janell Blackmer, and Ashley Johnson. 2021. Optimizing implementation in cancer control (opticc): protocol for an implementation science center. *Implementation Science Communications*, 2(1):1–16.

Cara C Lewis, Predrag Klasnja, Aaron R Lyon, Byron J Powell, Rebecca Lengnick-Hall, Gretchen Buchanan, Rosemary D Meza, Michelle C Chan, Marcella H Boynton, and Bryan J Weiner. 2022. The mechanics of implementation strategies and measures: advancing the study of implementation mechanisms. *Implementation Science Communications*, 3(1):1–11.

Cara C Lewis, Predrag Klasnja, Byron J Powell, Aaron R Lyon, Leah Tuzzio, Salene Jones, Callie Walsh-Bailey, and Bryan Weiner. 2018. From classification to causality: advancing understanding of mechanisms of change in implementation science. *Frontiers in public health*, 6:136.

Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What can you do? studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 264–275.

Hajin Lim. 2021. *Designing to Support Sensemaking in Cross-Lingual Computer-Mediated Communication Using NLP Techniques*. Ph.D. thesis, Cornell University.

Michelle P Lin, Bonnie B Blanchfield, Rose M Kakoza, Vineeta Vaidya, Christin Price, Joshua S Goldner, Michelle Higgins, Elisabeth Lessenich, Karl Laskowski, Jeremiah D Schuur, et al. 2017. Ed-based care coordination reduces costs for frequent ed users.

Ross James Lordon. 2019. *Design, Development, and Evaluation of a Patient-Centered Health Dialog System to Support Inguinal Hernia Surgery Patient Information-Seeking*. Ph.D. thesis, University of Washington.

Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51.

Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet interventions*, 10:39–46.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631.

Aaron R Lyon, Stephanie K Brewer, and Patricia A Areán. 2020. Leveraging human-centered design to implement modern psychological science: Return on an early investment. *American Psychologist*, 75(8):1067.

Aaron R Lyon and Eric J Bruns. 2019. User-centered redesign of evidence-based psychosocial interventions to enhance implementation—hospitable soil or better seeds? *JAMA psychiatry*, 76(1):3–4.

Aaron R Lyon, Sean A Munson, Brenna N Renn, David C Atkins, Michael D Pullmann, Emily Friedman, and Patricia A Areán. 2019. Use of human-centered design to improve implementation of evidence-based psychotherapies in low-resource communities: protocol for studies applying a framework to assess usability. *JMIR research protocols*, 8(10):e14990.

Martin Maguire. 2001. Methods to support human-centred design. *International journal of human-computer studies*, 55(4):587–634.

Patrick W Malecha, James H Williams, Nathan M Kunzler, Lewis R Goldfrank, Harrison J Alter, and Kelly M Doran. 2018. Material needs of emergency department patients: a systematic review. *Academic Emergency Medicine*, 25(3):330–359.

Jennifer Mankoff, Anind K Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. 2003. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 169–176. ACM.

Mark C Marino. 2014. The racial formation of chatbots. *CLCWeb: Comparative Literature and Culture*, 16(5):13.

Michael Marmot. 2005. Social determinants of health inequalities. *The lancet*, 365(9464):1099–1104.

Michael Marmot, Sharon Friel, Ruth Bell, Tanja AJ Houweling, and Sebastian Taylor. 2008. Closing the gap in a generation: health equity through action on the social determinants of health. *The lancet*, 372(9650):1661–1669.

Jessie Kimbrough Marshall, Olive M Mbah, Jean G Ford, Darcy Phelan-Emrick, Saifuddin Ahmed, Lee Bone, Jennifer Wenzel, Gary R Shapiro, Mollie Howerton, Lawrence Johnson, et al. 2016. Effect of patient navigation on breast cancer screening among african american medicare beneficiaries: a randomized controlled trial. *Journal of general internal medicine*, 31:68–76.

James C McCroskey, Virginia P Richmond, and John A Daly. 1975. The development of a measure of perceived homophily in interpersonal communication. *Human Communication Research*, 1(4):323–332.

Centers for Medicare and Medicaid Services. 2019. The accountable health communities health-related social needs screening tool.

Marijke Melles, Armagan Albayrak, and Richard Goossens. 2021. Innovating health care: key characteristics of human-centered design. *International Journal for Quality in Health Care*, 33(Supplement_1):37–44.

Sanjana Mendu, Mehdi Boukhechba, Janna R Gordon, Debajyoti Datta, Edwin Molina, Gloria Arroyo, Sara K Proctor, Kristen J Wells, and Laura E Barnes. 2018. Design of a culturally-informed virtual human for educating hispanic women about cervical cancer. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 360–366.

Susan Michie, Robert West, Kate Sheals, and Cristina A Godinho. 2018. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. *Translational behavioral medicine*, 8(2):212–224.

Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research*, 22(10):e20346.

David C Mohr, Aaron R Lyon, Emily G Lattie, Madhu Reddy, and Stephen M Schueller. 2017. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. *Journal of medical Internet research*, 19(5):e7725.

Rolf Molich and Jakob Nielsen. 1990. Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348.

Yamile Molina, Sage Kim, Nerida Berrios, and Elizabeth A Calhoun. 2015. Medical mistrust and patient satisfaction with mammography: the mediating effects of perceived self-efficacy among navigated african american women. *Health Expectations*, 18(6):2941–2950.

Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67.

Nancy S Morris, Charles D MacLean, Lisa D Chew, and Benjamin Littenberg. 2006. The single item literacy screener: evaluation of a brief instrument to identify limited reading ability. *BMC family practice*, 7:1–7.

Sarah Ann Mummah, Thomas N Robinson, Abby C King, Christopher D Gardner, and Stephen Sutton. 2016. Ideas (integrate, design, assess, and share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. *Journal of medical Internet research*, 18(12):e317.

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):33–45.

Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. *Digital health*, 5:2055207619871808.

NCCMT. 2023. Search | national collaborating centre for methods and tools.

Lisa A Newman and Linda M Kaljee. 2017. Health disparities and triple-negative breast cancer in african american women: a review. *JAMA surgery*, 152(5):485–493.

Hien Nguyen and Judith Masthoff. 2009. Designing empathic computers: the effect of multimodal empathic feedback using animated agent. In *Proceedings of the 4th international conference on persuasive technology*, pages 1–9.

Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *Knowledge and Systems Sciences: 18th International Symposium, KSS 2017, Bangkok, Thailand, November 17–19, 2017, Proceedings 18*, pages 38–52. Springer.

Jakob Nielsen. 1994. How to conduct a heuristic evaluation. `https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/`.

Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM.

Roobina Ohanian. 1990. Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of advertising*, 19(3):39–52.

Teresa K O'Leary, Elizabeth Stowell, Everlyne Kimani, Dhaval Parmar, Stefan Olafsson, Jessica Hoffman, Andrea G Parker, Michael K Paasche-Orlow, and Timothy Bickmore. 2020. Community-based cultural tailoring of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Bridget A Oppong, Samilia Obeng-Gyasi, Theresa Relation, and Lucile Adams-Campbell. 2021. Call to action: breast cancer screening recommendations for black women. *Breast Cancer Research and Treatment*, 187:295–297.

Chinelo C Orji, Chisom Kanu, Anuoluwapo I Adelodun, and Carolyn M Brown. 2020. Factors that influence mammography use for breast cancer screening among african american women. *Journal of the National Medical Association*, 112(6):578–592.

Alicia O'Cathain, Liz Croot, Katie Sworn, Edward Duncan, Nikki Rousseau, Katrina Turner, Lucy Yardley, and Pat Hoddinott. 2019. Taxonomy of approaches to developing interventions to improve health: a systematic methods overview. *Pilot and feasibility studies*, 5(1):1–27.

Karey L O'Hara, Lindsey M Knowles, Kate Guastaferro, and Aaron R Lyon. 2022. Human-centered design methods to achieve preparation phase goals in the multiphase optimization strategy framework. *Implementation Research and Practice*, 3:26334895221131052.

Laura Packel, Carolyn Fahey, Prosper Njau, and Sandra I McCoy. 2019. Implementation science using proctor's framework and an adaptation of the multiphase optimization strategy (most): Optimizing a financial incentive intervention for hiv treatment adherence in tanzania. *Journal of acquired immune deficiency syndromes (1999)*, 82(Suppl 3):S332.

Lawrence A Palinkas, Gregory A Aarons, Sarah Horwitz, Patricia Chamberlain, Michael Hurlburt, and John Landsverk. 2011. Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 38:44–53.

Victoria A Parker and Christy Harris Lemak. 2011. Navigating patient navigation: crossing health services research and clinical boundaries. *Biennial review of health care management*, pages 149–183.

Susan Racine Passmore, Kester F Williams-Parry, Erica Casper, and Stephen B Thomas. 2017. Message received: African american women and breast cancer screening. *Health promotion practice*, 18(5):726–733.

Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

Sabita Persaud. 2018. Addressing social determinants of health through advocacy. *Nursing Administration Quarterly*, 42(2):123–128.

Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. 2017. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports*, 7(1):1–7.

Byron J Powell, Rinad S Beidas, Cara C Lewis, Gregory A Aarons, J Curtis McMillen, Enola K Proctor, and David S Mandell. 2017. Methods to improve the selection and tailoring of implementation strategies. *The journal of behavioral health services & research*, 44:177–194.

Byron J Powell, Thomas J Waltz, Matthew J Chinman, Laura J Damschroder, Jeffrey L Smith, Monica M Matthieu, Enola K Proctor, and JoAnn E Kirchner. 2015. A refined compilation of implementation strategies: results from the expert recommendations for implementing change (eric) project. *Implementation science*, 10(1):1–14.

Alexandra Power-Hays, Stephanie Li, Akosua Mensah, and Amy Sobota. 2020. Universal screening for social determinants of health in pediatric sickle cell disease: a quality-improvement initiative. *Pediatric blood & cancer*, 67(1):e28006.

Enola Proctor, Hiie Silmere, Ramesh Raghavan, Peter Hovmand, Greg Aarons, Alicia Bunger, Richard Griffey, and Melissa Hensley. 2011. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and policy in mental health and mental health services research*, 38:65–76.

Enola K Proctor, Byron J Powell, and J Curtis McMillen. 2013. Implementation strategies: recommendations for specifying and reporting. *Implementation science*, 8(1):1–11.

Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6):785–797.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.

Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, et al. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint arXiv:1806.10698*.

Hyekyun Rhee, James Allen, Jennifer Mammen, and Mary Swift. 2014. Mobile phone-based asthma self-

management aid for adolescents (masmaa): a feasibility study. *Patient preference and adherence*, pages 63–72.

Raoul Rickenberg and Byron Reeves. 2000. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 49–56.

Arlinda Ruco, Fahima Dossa, Jill Tinmouth, Diego Llovet, Jenna Jacobson, Teruko Kishibe, and Nancy Baxter. 2021. Social media and mhealth technology for cancer screening: systematic review and meta-analysis. *Journal of Medical Internet Research*, 23(7):e26759.

Kathleen M Russell, Victoria L Champion, Patrick O Monahan, Sandra Millon-Underwood, Qianqian Zhao, Nicole Spacey, Nathan L Rush, and Electra D Paskett. 2010. Randomized trial of a lay health advisor and computer intervention to increase mammography screening in african american women. *Cancer Epidemiology, Biomarkers & Prevention*, 19(1):201–210.

Rachel C Shelton, Prajakta Adsul, April Oh, Nathalie Moise, and Derek M Griffith. 2021. Application of an antiracism lens in the field of implementation science (is): recommendations for reframing implementation research with a focus on justice and racial equity. *Implementation Research and Practice*, 2:26334895211049482.

Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2):47–53.

Priti Singh, Pallavi Jonnalagadda, Evan Morgan, and Naleef Fareed. 2022. Outpatient portal use in prenatal care: differential use by race, risk, and area social determinants of health. *Journal of the American Medical Informatics Association*, 29(2):364–371.

Rebecca Smith-Bindman, Diana L Miglioretti, Nicole Lurie, Linn Abraham, Rachel Ballard Barbash, Jodi Strzelczyk, Mark Dignan, William E Barlow, Cherry M Beasley, and Karla Kerlikowske. 2006. Does utilization of screening mammography explain racial and ethnic differences in breast cancer? *Annals of internal medicine*, 144(8):541–553.

Leslie B Snyder and Jessica M LaCroix. 2001. How effective are mediated health campaigns. *Public communication campaigns*, 3:181–190.

Elizabeth Stowell, Mercedes C Lyson, Herman Saksono, Reneé C Wurth, Holly Jimison, Misha Pavel, and Andrea G Parker. 2018. Designing and evaluating mhealth interventions for vulnerable populations: A systematic review. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Suela Sulo, Josh Feldstein, Jamie Partridge, Bjoern Schwander, Krishnan Sriram, and Wm Thomas Summerfelt. 2017. Budget impact of a comprehensive nutrition-focused quality improvement program for malnourished hospitalized patients. *American health & drug benefits*, 10(5):262.

Kumara Raja Sundar. 2018. Universal screening for social needs in a primary care clinic: a quality improvement approach using the your current life situation survey. *The Permanente Journal*, 22.

Nina Svenningsson and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pages 151–161.

Ming Tai-Seale, N Lance Downing, Veena Goel Jones, Richard V Milani, Beiqun Zhao, Brian Clay, Christopher Demuth Sharp, Albert Solomon Chan, and Christopher A Longhurst. 2019. Technology-enabled consumer engagement: promising practices at four health care delivery organizations. *Health Affairs*, 38(3):383–390.

Paula Tanabe, Rick Gimbel, Paul R Yarnold, Demetrios N Kyriacou, and James G Adams. 2004. Reliability and validity of scores on the emergency severity index version 3. *Academic emergency medicine*, 11(1):59–65.

Silke Ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design features of embodied conversational agents in ehealth: a literature review. *International Journal of Human-Computer Studies*, 138:102409.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in informa-

tion seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1171–1180.

Hayley S Thompson, Heiddis B Valdimarsdottir, Gary Winkel, Lina Jandorf, and William Redd. 2004. The group-based medical mistrust scale: psychometric properties and association with breast cancer screening. *Preventive medicine*, 38(2):209–218.

Sebastian T Tong, Winston R Liaw, Paulette Lail Kashiri, James Pecsok, Julia Rozman, Andrew W Bazemore, and Alex H Krist. 2018. Clinician experiences with screening for social needs in primary care. *The Journal of the American Board of Family Medicine*, 31(3):351–363.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158.

Kim M Unertl, Chris L Schaefbauer, Terrance R Campbell, Charles Senteio, Katie A Siek, Suzanne Bakken, and Tiffany C Veinot. 2016. Integrating community-based participatory research and informatics approaches to improve the engagement and health of underserved populations. *Journal of the American Medical Informatics Association*, 23(1):60–73.

Pablo Buitron de la Vega, Stephanie Losi, Linda Sprague Martinez, Allison Bovell-Ammon, Arvin Garg, Thea James, Alana M Ewen, Marna Stack, Heloisa DeCarvalho, Megan Sandel, et al. 2019. Implementing an ehr-based screening and referral system to address social determinants of health in primary care. *Medical care*, 57:S133–S139.

Tiffany C Veinot, Terrance R Campbell, Daniel J Kruger, and Alison Grodzinski. 2013. A question of trust: user-centered design requirements for an informatics intervention to promote the sexual health of african-american youth. *Journal of the American Medical Informatics Association*, 20(4):758–765.

Serena Vita, Raffaella Marocco, Irene Pozzetto, Giuseppe Morlino, Ester Vigilante, Vania Palmacci, Laura Fondaco, Blerta Kertusha, Monica Renzelli, Vito Mercurio, et al. 2018. The'doctor apollo'chatbot: a digital health tool to improve engagement of people living with hiv. *J Int AIDS Soc*, 21(Suppl 8):e25187.

Andrea S Wallace, Brenda Luther, Jia-Wen Guo, Ching-Yu Wang, Shawna Sisler, and Bob Wong. 2020. Implementing a social determinants screening and referral infrastructure during routine emergency department visits, utah, 2017–2018. *Preventing chronic disease*, 17:E45.

Haolin Wang, Qingpeng Zhang, Mary Ip, and Joseph Tak Fai Lau. 2018. Social media–based conversational agents for health management and interventions. *Computer*, 51(8):26–33.

Zhuxiaona Wei and James A Landay. 2018. Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Computing*, 17(2):84–96.

Bryan J Weiner, Cara C Lewis, Cameo Stanick, Byron J Powell, Caitlin N Dorsey, Alecia S Clary, Marcella H Boynton, and Heather Halko. 2017. Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science*, 12:1–12.

Delia Smith West, Paul Greene, LeaVonne Pulley, Polly Kratt, Stacy Gore, Heidi Weiss, and Nicole Siegfried. 2004. Stepped-care, community clinic interventions to promote mammography use among low-income rural african american women. *Health education & behavior*, 31(4_suppl):29S–44S.

Diahann Wilcox, Paula S McCauley, Colleen Delaney, and Sheila L Molony. 2018. Evaluation of a hospital: community partnership to reduce 30-day readmissions. *Professional case management*, 23(6):327–341.

Nick Wilson, E Jane MacDonald, Osman David Mansoor, and Jane Morgan. 2017. In bed with siri and google assistant: a comparison of sexual health advice. *Bmj*, 359.

K Witte, E Maibach, and RL Parrot. 1995. Using the persuasive health message framework to generate effective campaign messages. *Designing health messages*, pages 145–164.

Steven H Woolf, Richard Grol, Allen Hutchinson, Martin Eccles, and Jeremy Grimshaw. 1999. Potential benefits, limitations, and harms of clinical guidelines. *Bmj*, 318(7182):527–530.

Albert W Wu, Christine M Weston, Chidinma A Ibe, Claire F Ruberman, Lee Bone, Romsai T Boonyasai, Sandra Hwang, Janice Gentry, Leon Purnell, Yanyan Lu, et al. 2019. The baltimore community-based organizations neighborhood network: enhancing capacity together (connect) cluster rct. *American journal of preventive medicine*, 57(2):e31–e41.

Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental workload and language production in non-native speaker ipa interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(3):1–37.

Rosalie F Young, Kendra Schwartz, and Jason Booza. 2011. Medical barriers to mammography screening of african american women in a high cancer mortality area: implications for cancer educators and health providers. *Journal of Cancer Education*, 26:262–269.

Jiajie Zhang, Todd R Johnson, Vimla L Patel, Danielle L Paige, and Tate Kubose. 2003. Using usability heuristics to evaluate patient safety of medical devices. *Journal of biomedical informatics*, 36(1-2):23–30.

Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(2-3):1–36.

Kangmin Zhu, Sandra Hunter, Louis J Bernard, Kathleen Payne-Wilks, Chanel L Roland, Lloyd C Elam, Ziding Feng, and Robert S Levine. 2002. An intervention study on screening for breast cancer among single african-american women aged 65 and older. *Preventive Medicine*, 34(5):536–545.

Donna M Zulman, Matthew L Maciejewski, Janet M Grubber, Hollis J Weidenbacher, Dan V Blalock, Leah L Zullig, Liberty Greene, Heather E Whitson, Susan N Hastings, and Valerie A Smith. 2020. Patient-reported social and behavioral determinants of health and estimated risk of hospitalization in high-risk veterans affairs patients. *JAMA Network Open*, 3(10):e2021457–e2021457.

Chapter A

Appendix

A.1 Chapter 3: Heuristic Evaluation

A.1.1 Phase 3: In-person Heuristic Evaluation Instructions

Part I: Please read and familiarize yourself with the list of conversational agent-specific heuristics provided by the study administrator. This set of heuristics has been developed to describe common properties of usable conversational agents. You can refer back to this list as you examine the conversational agent.

Part II: You will be asked to evaluate an Alexa skill using the Amazon Echo. Please read the following description of the conversational agent that you will evaluate: "Slack With Voice is an unofficial skill that connects your Slack workspace with Amazon Echo. The Alexa skill can be used to send, read and react to messages on your Slack workspace." We have set up a fictional Slack workspace that has been linked to the Amazon Echo. The workspace is titled "Department of Human Centered Design & Engineering" and you will be using the username "Anna" to communicate with other people and classmates in the department.

Part III: We ask that you examine the conversational agent interface at least twice. In the first pass, spend 10-15 minutes to examine the interface to understand the flow of the interaction and the scope of the conversational agent. In the second pass, move through and analyze the interface against the defined principles (“heuristics”). When you identify an issue or area for improvement, record it in the table below with reference to one or more of the heuristics.

A.1.2 Phase 3: Online Heuristic Evaluation Instructions

Part I: Please follow the link below and familiarize yourself with the set of heuristics. This set of heuristics has been developed to describe common properties of usable conversational agents. Please refer back to this list as you examine the conversational agent.

Part II: Please read the following description of the conversational agent that you will evaluate: "Harbor Bot is a text-based conversational agent that is designed to collect survey information in hospital emergency departments. Harbor Bot asks users various questions regarding their health, housing situation, and employment, to screen users for unmet social needs." Follow this link to access the conversational agent that you will be evaluating.

Part III: We ask that you examine the conversational agent interface at least twice. In the first pass, examine the interface to understand the flow of the interaction and the scope of the conversational agent. In the second pass, move through and analyze the interface against the defined principles ("heuristics"). When you identify an issue or area for improvement, record it in the table below with reference to one or more of the heuristics.

A.1.3 Phase 4: Online Heuristic Evaluation Instructions

Part I: Please follow the link below and familiarize yourself with the set of heuristics that you will use for the heuristic evaluation. This set of heuristics has been developed to describe general principles for the visual and interaction design of conversational user interfaces. These are 11 general principles for the visual and interaction design of conversational user interfaces (including graphical user interfaces, voice user interfaces, and multimodal interfaces). They are called "heuristics" because they are more in the nature of rules of thumb than specific usability guidelines. Please refer back to this list as you examine the conversational agent.

Part II: Please read the following description of the conversational agent that you will evaluate: "Harbor Bot is a text-based conversational agent that is designed to collect survey information in hospital emergency departments. Harbor Bot asks users various questions regarding their health, housing situation, and employment, to screen users for unmet social needs." Follow this link to access the conversational agent that you will be evaluating.

Part III: We ask that you examine the conversational user interface at least twice. In the first pass, examine the interface to understand the flow of the interaction and the scope of the conversational agent. In the second pass, move through and analyze the interface against the defined principles ("heuristics"). When you identify an issue, record it in the table below with reference to one or more of the heuristics.

A.2 Chapter 4: Screening Questionnaire

Social Needs Questions

1. Could I ask you, do you have the resources to pay for the very basics like food, housing, medical care, and heating?
2. Further, do you have any significant outstanding bills or debts?
3. Are you currently receiving Social Security benefits?
4. Do you expect to be out of work for at least 12 months?
5. Do you want help finding or keeping work or a job?
6. Continuing, do you want help with school or training?
7. Think about the place you live. Do you have any problems with the following? Choose all that apply.
8. What is your living situation today?
9. In the past 12 months, has the electric, gas, oil or water company threatened to shut off services in your home?
10. Within the past 12 months, were you worried whether your food would run out before you got money to buy more?
11. Within the past 12 months, the food you bought just didn't last and you didn't have money to get more?
12. Next, could you tell me whether in the past 12 months, has lack of reliable transportation kept you from medical appointments, meetings, work or from getting things needed for daily living?

Phase 1	Rel.	Phase 2
Visibility of system status The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. The system should allow the user to request information or identify what is occurring.	3.7	**Visibility of system status** The system should always keep users informed about what is going on, through appropriate feedback within reasonable time, without overwhelming the user. The user should be allowed to request information about the system status.
Clarify capabilities Ensure users get a sense of system capabilities by using clarifications throughout the conversational agent use. The system should also clearly indicate that it is not a human.	4	**Clarify capabilities** Ensure users get a sense of system capabilities through appropriate design and clarifications (either implicitly or explicitly) through the conversational agent interaction. The system should not falsely claim to be a human.
Match between system and the real world The system should understand and speak the users' language—with words, phrases and concepts familiar to the user—rather than system-oriented terms or confusing terminology. Mirror real life conversations and include dialogue elements that create a smooth conversation through openings, mid-conversation guidance, and graceful exits. In domains that are focused on functional support, rather than emotional support, limit social-based characteristics.	4.1	**Match between system and the real world** The system should understand and speak the users' language—with words, phrases and concepts familiar to the user—rather than system-oriented terms or confusing terminology. Include dialogue elements that create a smooth conversation through openings, mid-conversation guidance, and graceful exits.
User control and freedom Users often choose system functions by mistake and will need an option to effortlessly leave the unwanted state without having to go through an extended dialogue. Support undo and redo, and allow users to control the repair of errors.	4	**User control and freedom** Some system functions may be chosen by mistake and will need an option to effortlessly leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
Consistency and standards Users should not have to wonder whether different words, situations, or actions mean the same thing across contexts of use. Within the interaction, the system should have a consistent voice, style of language, and personality. Users should be able to receive consistent responses even if they communicate the same function in multiple ways.	4.3	**Consistency and standards** Users should not have to wonder whether different words, situations, or actions mean the same thing. Users should also be able to receive consistent responses even if they communicate the same function in multiple ways (and modalities). Within the interaction, the system should have a consistent voice, style of language, and personality.
Error prevention Even better than good error messages is a careful design of the conversation and interface to reduce the likelihood of a problem from occurring in the first place. Be prepared for dialogue failures, deadends or sidetracks. Either proactively prevent or eliminate potential error-prone conditions, or check and confirm with users before they commit an action.	3.9	**Error prevention** Even better than good error messages is a careful design of the conversation and interface to reduce the likelihood of a problem from occurring in the first place. Be prepared for dialogue failures, deadends or sidetracks. Proactively prevent or eliminate potential error-prone conditions, and check and confirm with users before they commit an action.

Table A.1: The 12 conversational agent heuristics compared to the earlier modified heuristics, and the average relevance rating for each heuristic.

Phase 1	Rel.	Phase 2
Recognition rather than recall Minimize the user's memory load by making objects, actions, and options clear to users. The system should minimize the information remembered from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.	3.8	**Learnability** Minimize the user's cognitive load by guiding and prompting the users (either implicitly or explicitly) throughout the dialogue. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
Domain specific flexibility and efficiency of use Provide domain specific enhanced functionalities and accelerators to ensure that the system is useful and efficient compared to existing alternatives. Allow users the ability to interact with the system using the appropriate or their preferred modality and hardware.	3.8	**Multimodal flexibility and efficiency of use** Support flexible interactions by allowing users to interact with the system using appropriate and/or preferred modality and hardware. Additionally, provide accelerators, such as verbal shortcuts that are unseen by novices but speed up the interactions for experts, to ensure that the system is efficient.
Aesthetic, minimalist and engaging design Dialogues should not contain information which is irrelevant or rarely needed. Only provide interactional elements that are necessary to engage the user and fit within the goal of the system. Voice interfaces should support short interactions and expand on the conversation if the user chooses.	4.1	**Aesthetic, minimalist and engaging design** Dialogues should not contain information which is irrelevant or rarely needed. Provide interactional elements that are necessary to engage the user and fit within the goal of the system. Voice interfaces should support short interactions and expand on the conversation if the user chooses.
	N/A	**Help users recognize, diagnose and recover from errors** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
Help and documentation The system should provide help and documentation regarding the system's capabilities and script. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.	2.7	
Context preservation The system should maintain context preservation regarding the conversation topic, intra- and inter-session. Allow the user to reference past messages for further interactions to support implicit user expectations of conversations.	4	**Context preservation** Maintain context preservation regarding the conversation topic intra-session, and if possible inter-session. Allow the user to reference past messages for further interactions to support implicit user expectations of conversations.
Privacy The system should convey trustworthiness and reliability by providing the user with information about the privacy of their data.	4.1	**Trustworthiness** The system should convey trustworthiness by ensuring privacy of user data, and by being transparent and truthful with the user.
Veracity Be honest with the user by providing accurate information within the dialogue.	3.8	

Table A.2: The 12 conversational agent heuristics compared to the earlier modified heuristics, and the average relevance rating for each heuristic (cont.)

13. Are you or anyone in your household having issues with any of the following?

14. Could you tell me in the past year, how often have you used alcohol?

15. In the past year, how often have you used prescription drugs for non-medical reasons?

16. In the past year, how often have you used illegal drugs?

Demographic Questions

1. What is your racial and ethnic background? Check all that apply.

2. What is your gender identity? Check all that apply.

3. What is your age? Please press Return or Enter when you have finished typing.

4. What type of health insurance do you have?

5. What is the highest degree or level of school you have completed?

6. What is your current marital or relationship status?

Implementation Outcome Measures

To what extend do you agree with the following statements:

1. I like the use of this chatbot to answer these questions.

2. Using this chatbot to answer these questions seems suitable.

3. Using this chatbot to answer these questions seems easy to use.

A.3 Chapter 4: Interview Guide

I'd like to ask you some questions about your recent visit to the ED where you used a chatbot to answer questions about social and material needs. In your responses, you indicated that you would be willing to be contacted and interviewed about your experience. We will provide you with $30 compensation for your

time. In this interview, I will ask about your experience and how well the chatbot worked for you. Do you confirm your consent to be interviewed? For the purpose of this research, may I record this conversation?

Description of Workflow Experience

- I'd like to ask you about your experience in the ED waiting room before being seen by a medical provider.
- What was the primary reason for your ED visit?
- Do you remember sitting in the waiting room? What was that experience like?
- Do you remember being approached about using the chatbot?

I'm now going to ask you some questions that are specific to the screening we did in ED and your experience related to that.

Implementation Outcomes: Acceptability

- Do you believe it was acceptable to be asked about social and material needs (such as food and housing) during your emergency department visit?
- Did you have any reservations or concerns about being asked about your social and material needs (such as food and housing) during your emergency department visit?
- Would you be willing to answer questions about social and material needs (such as food and housing) during future ED visits? In other clinical settings?

Thinking about the chatbot tool we used to ask about social and material needs (such as food and housing), please answer the following questions:

- From 1 to 5, 1 being completely disagree and 5 being completely agree: I like using the chatbot to answer questions about social and material needs).
- Thinking about your experience answering questions related to your social and material needs (such as food and housing), did you think the chatbot was an acceptable way of answering these questions?

  - What did you like about using the chatbot to answer these questions?
  - What did you dislike about using the chatbot to answer these questions?

- About how long did it take for you to answer the questions on the chatbot? Was this amount of time acceptable to you?

Implementation Outcomes: Feasibility / Usability

- From 1 to 5, 1 being completely disagree and 5 being completely agree: Using the chatbot to answer questions about social and material seems suitable.
- Was the chatbot easy to use? Did you have any difficulties using the chatbot?
- Were there any questions that were difficult to understand? Were there any questions that were difficult to answer? Were the answer options appropriate?
- Would you be willing to answer questions about social and material needs using the chatbot during future ED visits? In other clinical settings?

Implementation Outcomes: Appropriateness

- From 1 to 5, 1 being completely disagree and 5 being completely agree: Using the chatbot to answer questions about social and material seems easy to use.
- Why (or why not) was it appropriate to use the chatbot to answer these questions?
- Do you have any concerns about answering any of the questions on the chatbot?
- Was it appropriate to use the chatbot in the ED waiting room? If not, why?

General Feedback

- Is there anything you would change about the chatbot?

Resource Usage

- At the end of the chatbot screening, you were given a handout with a list of resources. Did you discuss your responses to the social and material needs screener with your doctor, nurse or anyone else in the ED? What did you talk about?
  - If you shared this information with your doctor, were you comfortable doing so? If not, why not?
  - Have you shared information like this with a doctor before?
  - Would you be interested in having your responses to the social and material needs screener shared with social service providers in the community?
- Did you try to contact any of the social service providers?
  - If yes, were you able to speak with anyone? (if not) Why not?
  - Did you visit them in person? (if not) Why not?
  - If no, can you tell us why you chose not to reach out to these providers?
- Were there any resources on the list you were not aware of before you received the handout?
- Have you contacted social service providers before?
  - Did you have any problems finding or connecting to them in the past?
  - Were they helpful?

Administer Survey: Health Literacy

I want to ask you a quick question about your general understanding of health-related materials.

- How often do you need to have someone help you when you read instructions, pamphlets, or other written material from your doctor or pharmacy? (SILS): 1-Never, 2-Rarely, 3-Sometimes, 4-Often, and 5-Always

Wrap Up

- Is there anything else you would like to share with us?
- Is this phone number okay for sending you a gift card?

Table A.3: Response distributions of acceptability ratings by age, ethnicity, and education

	Acceptability	Completely disagree (%)	Disagree (%)	Neither agree nor disagree (%)	Agree (%)	Completely agree (%)	Average rating on 1-5 Likert scale (SD)
	All participants	13 (4.4)	8 (2.7)	53 (17.9)	136 (45.8)	87 (29.3)	3.93 (0.98)
Age (y)	18-25	0 (0.0)	0 (0.0)	3 (11.1)	16 (59.3)	8 (29.6)	4.19 (0.61)
	26-35	4 (5.2)	2 (2.6)	16 (20.8)	35 (45.5)	20 (26.0)	3.84 (1.01)
	36-45	1 (1.9)	4 (7.7)	8 (15.4)	24 (46.2)	15 (28.9)	3.92 (0.96)
	46-55	3 (9.7)	0 (0.0)	4 (12.9)	11 (35.5)	13 (41.9)	4.0 (1.19)
	56-65	1 (5.0)	2 (10.0)	2 (10.0)	11 (55.0)	4 (20.0)	3.75 (1.04)
	>66	1 (5.0)	0 (0.0)	6 (30.0)	7 (35.0)	6 (30.0)	3.85 (1.01)
	Prefer not to answer	3 (4.3)	0 (0.0)	14 (20.0)	32 (45.7)	21 (30.0)	3.97 (0.94)
Racial/Ethnic Background	White	3 (2.5)	1 (0.8)	24 (20.2)	57 (47.9)	34 (28.6)	3.99 (0.86)
	Black, African American or African	5 (7.8)	1 (1.6)	7 (10.9)	29 (45.3)	22 (34.4)	3.97 (1.10)
	Latin American, Central American, Mexican or Mexican American, Hispanic or Chicano	3 (5.9)	2 (3.9)	7 (13.7)	25 (49.0)	14 (27.5)	3.88 (1.04)
	More than one race	0 (0.0)	1 (3.3)	6 (20.0)	14 (46.7)	9 (30.0)	4.03 (0.80)
	Asian: Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other	1 (7.7)	2 (15.4)	3 (23.1)	2 (15.4)	5 (38.5)	3.62 (1.33)
	Other	1 (7.7)	1 (7.7)	3 (23.1)	6 (46.2)	2 (15.4)	3.54 (1.08)
	Prefer not to answer	0 (0.0)	0 (0.0)	3 (42.9)	3 (42.9)	1 (14.3)	3.71 (0.70)
Education	Some college	2 (2.7)	0 (0.0)	13 (17.8)	38 (52.1)	20 (27.4)	4.01 (0.84)
	High school graduate	3 (4.6)	2 (3.0)	10 (15.2)	32 (48.5)	19 (28.8)	3.94 (0.98)
	Bachelor's degree	1 (2.6)	0 (0.0)	11 (29.0)	17 (44.7)	9 (23.7)	3.87 (0.86)
	Less than high school	3 (9.4)	0 (0.0)	4 (12.5)	11 (34.4)	14 (43.8)	4.03 (1.19)
	Some high school	4 (14.8)	3 (11.1)	2 (7.4)	10 (37.0)	8 (29.6)	3.56 (1.40)
	Graduate school	0 (0.0)	2 (7.4)	7 (25.9)	11 (40.7)	7 (25.9)	3.85 (0.89)
	Associate degree	0 (0.0)	1 (3.6)	6 (21.4)	12 (42.9)	9 (32.1)	4.04 (0.82)
	Prefer not to answer	0 (0.0)	0 (0.0)	0 (0.0)	5 (83.3)	1 (16.7)	4.17 (0.37)

Table A.4: Response distributions of feasibility ratings by age, ethnicity, and education

	Feasibility	Completely disagree (%)	Disagree (%)	Neither agree nor disagree (%)	Agree (%)	Completely agree (%)	Average rating on 1-5 Likert scale (SD)
	All participants	10 (3.3)	4 (1.3)	16 (5.3)	158 (52.5)	113 (37.5)	4.20 (0.86)
Age (y)	18-25	0 (0.0)	0 (0.0)	1 (3.7)	14 (51.9)	12 (44.4)	4.41 (0.56)
	26-35	1 (1.3)	0 (0.0)	3 (3.8)	45 (57.0)	30 (38.0)	4.30 (0.66)
	36-45	2 (3.8)	1 (1.9)	2 (3.8)	26 (49.1)	22 (41.5)	4.23 (0.90)
	46-55	0 (0.0)	0 (0.0)	2 (6.3)	15 (46.9)	15 (46.9)	4.41 (0.61)
	56-65	1 (5.6)	1 (5.6)	1 (5.6)	10 (55.6)	5 (27.8)	3.94 (1.03)
	>66	1 (5.0)	1 (5.0)	1 (5.0)	10 (50.0)	7 (35.0)	4.05 (1.02)
	Prefer not to answer	5 (6.9)	1 (1.4)	6 (8.3)	38 (52.8)	22 (30.6)	3.99 (1.03)
Racial/Ethnic Background	White	2 (1.7)	2 (1.7)	8 (6.7)	64 (53.3)	44 (36.7)	4.22 (0.78)
	Black, African American or African	5 (7.7)	0 (0.0)	3 (4.6)	29 (44.6)	28 (43.1)	4.15 (1.07)
	Latin American, Central American, Mexican or Mexican American, Hispanic or Chicano	2 (4.0)	1 (2.0)	1 (2.0)	28 (56.0)	18 (36.0)	4.18 (0.89)
	More than one race	0 (0.0)	1 (3.2)	2 (6.5)	15 (48.4)	13 (41.9)	4.29 (0.73)
	Asian: Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other	0 (0.0)	0 (0.0)	0 (0.0)	8 (61.5)	5 (38.5)	4.38 (0.49)
	Other	1 (7.7)	0 (0.0)	2 (15.4)	7 (53.9)	3 (23.1)	3.85 (1.03)
	Prefer not to answer	0 (0.0)	0 (0.0)	0 (0.0)	7 (77.8)	2 (22.2)	4.22 (0.42)
Education	Some college	1 (1.3)	0 (0.0)	3 (4.0)	42 (56.0)	29 (38.7)	4.31 (0.67)
	High school graduate	4 (6.0)	0 (0.0)	4 (6.0)	34 (50.8)	25 (37.3)	4.13 (0.98)
	Bachelor's degree	0 (0.0)	2 (5.3)	2 (5.3)	19 (50.0)	15 (39.5)	4.24 (0.78)
	Less than high school	0 (0.0)	0 (0.0)	3 (9.4)	13 (40.6)	16 (50.0)	4.41 (0.65)
	Some high school	1 (3.6)	2 (7.1)	0 (0.0)	16 (57.1)	9 (32.1)	4.07 (0.96)
	Graduate school	2 (7.1)	0 (0.0)	2 (7.1)	15 (53.6)	9 (32.1)	4.04 (1.02)
	Associate degree	0 (0.0)	0 (0.0)	2 (7.4)	15 (55.6)	10 (37.0)	4.30 (0.60)
	Prefer not to answer	2 (33.3)	0 (0.0)	0 (0.0)	4 (66.7)	0 (0.0)	3.0 (1.41)

Table A.5: Response distributions of appropriateness ratings by age, ethnicity, and education

	Appropriateness	Completely disagree (%)	Disagree (%)	Neither agree nor disagree (%)	Agree (%)	Completely agree (%)	Average rating on 1-5 Likert scale (SD)
	All participants	6 (2.0)	9 (3.0)	35 (11.6)	150 (49.7)	102 (33.8)	4.10 (0.86)
Age (y)	18-25	0 (0.0)	0 (0.0)	4 (14.8)	15 (55.6)	8 (29.6)	4.15 (0.65)
	26-35	1 (1.3)	1 (1.3)	8 (10.3)	45 (57.7)	23 (29.5)	4.13 (0.74)
	36-45	2 (3.6)	2 (3.6)	4 (7.3)	26 (47.3)	21 (38.2)	4.13 (0.95)
	46-55	0 (0.0)	2 (6.5)	2 (6.5)	14 (45.2)	13 (41.9)	4.23 (0.83)
	56-65	0 (0.0)	2 (10.5)	0 (0.0)	12 (63.2)	5 (26.3)	4.05 (0.83)
	>66	0 (0.0)	1 (5.0)	5 (25.0)	5 (25.0)	9 (45.0)	4.10 (0.94)
	Prefer not to answer	3 (4.2)	1 (1.4)	12 (16.7)	33 (45.8)	23 (31.9)	4.0 (0.96)
Racial/Ethnic Background	White	1 (0.9)	2 (1.7)	15 (12.7)	56 (47.5)	44 (37.3)	4.19 (0.78)
	Black, African American or African	3 (4.6)	4 (6.2)	4 (6.2)	31 (47.7)	23 (35.4)	4.03 (1.04)
	Latin American, Central American, Mexican or Mexican American, Hispanic or Chicano	1 (1.9)	2 (3.9)	5 (9.6)	29 (55.8)	15 (28.9)	4.06 (0.84)
	More than one race	0 (0.0)	1 (3.1)	3 (9.4)	15 (46.9)	13 (40.6)	4.25 (0.75)
	Asian: Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other	0 (0.0)	0 (0.0)	3 (23.1)	5 (38.5)	5 (38.5)	4.15 (0.77)
	Other	1 (7.1)	0 (0.0)	4 (28.6)	8 (57.1)	1 (7.1)	3.57 (0.90)
	Prefer not to answer	0 (0.0)	0 (0.0)	1 (12.5)	6 (75.0)	1 (12.5)	4.0 (0.50)
Education	Some college	2 (2.7)	2 (2.7)	5 (6.7)	38 (50.7)	28 (37.3)	4.17 (0.87)
	High school graduate	2 (3.0)	1 (1.5)	5 (7.6)	39 (59.1)	19 (28.8)	4.09 (0.83)
	Bachelor's degree	0 (0.0)	1 (2.6)	8 (21.1)	19 (50.0)	10 (26.3)	4.0 (0.76)
	Less than high school	0 (0.0)	1 (3.0)	5 (15.2)	10 (30.3)	17 (51.5)	4.30 (0.83)
	Some high school	1 (3.5)	2 (6.9)	1 (3.5)	17 (58.6)	8 (27.6)	4.0 (0.95)
	Graduate school	1 (3.6)	0 (0.0)	6 (21.4)	11 (39.3)	10 (35.7)	4.04 (0.94)
	Associate degree	0 (0.0)	1 (3.7)	4 (14.8)	12 (44.4)	10 (37.0)	4.15 (0.80)
	Prefer not to answer	0 (0.0)	1 (16.7)	1 (16.7)	4 (66.7)	0 (0.0)	3.50 (0.76)

A.4 Chapter 5: Interview Guide

- What do you know about breast cancer screening (mammograms)?
- Do you have a primary care provider?
  - If yes, has your PCP talked to you about breast cancer screening? Has anyone else from your primary care clinic talked to you about breast cancer screening? How was that experience?
- Who else have you discussed breast cancer screening with (other than health care professionals)?
- Have you had breast cancer screening (mammogram)?
  - If yes:
    * Why did you have breast cancer screening?
    * Where did you have breast cancer screening? (clinic, hospital, mobile van?)
    * Did anything make it easier or harder to get breast cancer screening?
  - If no:
    * Why haven't you had breast cancer screening?
    * Have you ever been recommended to have breast cancer screening?
    * If you were to get breast cancer screening, where would you get it and when (what time, day of week)?
- Have you had challenges with:
  - Scheduling appointments
  - Locating screening sites
  - Knowing whether and when you should get a mammogram
  - Other challenges?
- Has the COVID-19 pandemic prevented you from getting screening or affected how you think about screening?

- What else would you like to share that we haven't covered yet?

Our design team is working on an app to connect patients to more information about breast cancer screening. I'd like to show you a few screens to get your feedback.

1. Intro screen - chatbot where users can ask some questions about screening.
    - What kinds of questions would you like to ask the chatbot?
2. Scheduling screen - one of the key functions would be to help patients find screening locations.
    - Would a feature like this be useful (it might tie in with 4b above). What other features would you like? How would these features work? (What questions are missing?)
3. Barriers screen - another key function would be to use the chatbot to address some common concerns patients may have.
    - What do you like about this app?
    - What do you not like about this app?
    - Would you use this app to ask these questions? When would you use this?
    - What would you change about this app, if anything?

A.5 Chapter 5: Focus Group Guide

In these screens, the patient receives a message from Sesi that they are due for a mammogram. The patient asks for more information and Sesi guides them through questions about what to expect, etc. These screens include question answer videos.

Motivation to Use

- Does this encourage you to schedule a mammogram?
- Does the conversation feel engaging? If not, what additions would make you interested in continuing to participate with the app?

- Is the sentence about self-exams confusing? Does it help you understand the difference between self-exams and mammograms?
- Does seeing the image of the mammogram machine make you more or less comfortable with proceeding?

Perception of Chatbot Persona

- Would you feel comfortable to talk with Sesi about issues related to breast cancer screening?
- How much do you trust Sesi?
- Do you think you will use some of the information provided by Sesi?
- How relatable was Sesi? What was your first impression of Sesi?

In these screens, the patient is opening the app and is using it to schedule a mammogram appointment.

Desired Features

- Would you prefer to have an option in the app to request a female mammography technician?
- Are there other types of information that you would like from this chatbot?
- What else would you like to use this chatbot for?

Usability

- Based on the interactions that were shown, do you think that the system is easy to use?
- Do you foresee any difficulties while interacting with Sesi?
- Was there anything that Sesi said that was confusing?

In these screens, a patient previously scheduled an appointment, but they could not make their appointment. Sesi sends them a message to ask why they couldn't make an appointment and helps them to reschedule.

Motivation to Reschedule

- Would this encourage you to reschedule your appointment?

A.6 Chapter 5: Factorial Design Questionnaire

Thank you for your interest in this study. We are a team of researchers working to build a tool to support Black/African American women in receiving breast cancer screening.

Specifically, we are designing a chatbot which is a virtual guide that is an alternative to talking with a person. An example of a chatbot is the iPhone Siri or Amazon Alexa. The chatbot will be designed to share information about breast cancer screening and help with scheduling appointments.

The purpose of this study is to evaluate the first few messages for the chatbot. In this study, you will read and answer questions about the first few messages from the chatbot.

Please continue if you are a Black/African American woman and between the ages of 40-75 years old.

Please watch the video below and carefully read the first few messages sent by the chatbot. Once you have finished reading, please select 'Next'.

Please answer the following questions about the chatbot.

Intention to Use

1. What is the likelihood that you would use this chatbot to schedule a mammogram in the future?: 1-Very unlikely, 2-Unlikely, 3-Neutral, 4-Likely, and 5-Very likely

Engagement

Please rate the chatbot on the following trait:

1. Unimportant–important
2. Boring–interesting
3. Irrelevant–relevant
4. Cold-warm

Trust

To what extent do you agree with the following statements:

1. I believe that the chatbot will act in my best interest.
2. I believe that the chatbot is interested in understanding my needs and preferences.
3. I think that the chatbot is competent and effective in breast cancer screening education and scheduling.
4. I think that the chatbot performs its role in breast cancer screening education and scheduling very well.
5. I can trust the information presented to me by the chatbot.
6. I feel I must be cautious when using the chatbot.
7. It is risky to interact with the chatbot.

Directness Scales

Please rate the chatbot on the following trait:

1. Indirect-Direct
2. Unfriendly-Friendly
3. Unsympathetic-Caring
4. Ambiguous-Straightforward
5. Undemanding-Demanding
6. Disrespectful-Respectful
7. Impolite-Polite

Homophily

To what extent do you agree with the following statements:

1. The chatbot behaves like me.
2. The chatbot is similar to me.

Expertise

1. The chatbot is an expert.
2. The chatbot is knowledgeable.

Self-brand connection

1. The chatbot reflects who I am as a Black/African American woman.
2. I can identify with the chatbot as a Black/African American woman.
3. I feel a personal connection to the chatbot as a Black/African American woman.

Comfort with App Use Conversations about Breast Cancer

1. In general, how comfortable are you using chatbots to communicate?: 1-Very uncomfortable, 2-Uncomfortable, 3-Neutral, 4-Comfortable, 5-Very comfortable
2. In general, how comfortable are you talking about breast cancer with others?: 1-Very uncomfortable, 2-Uncomfortable, 3-Neutral, 4-Comfortable, 5-Very comfortable

Perception of the Persona (open-ended response)

1. How did you like or dislike the way the chatbot was presented?

Demographics

1. What is your age?

2. What is your zipcode?

3. If you are paying attention, please enter the number ’3’.

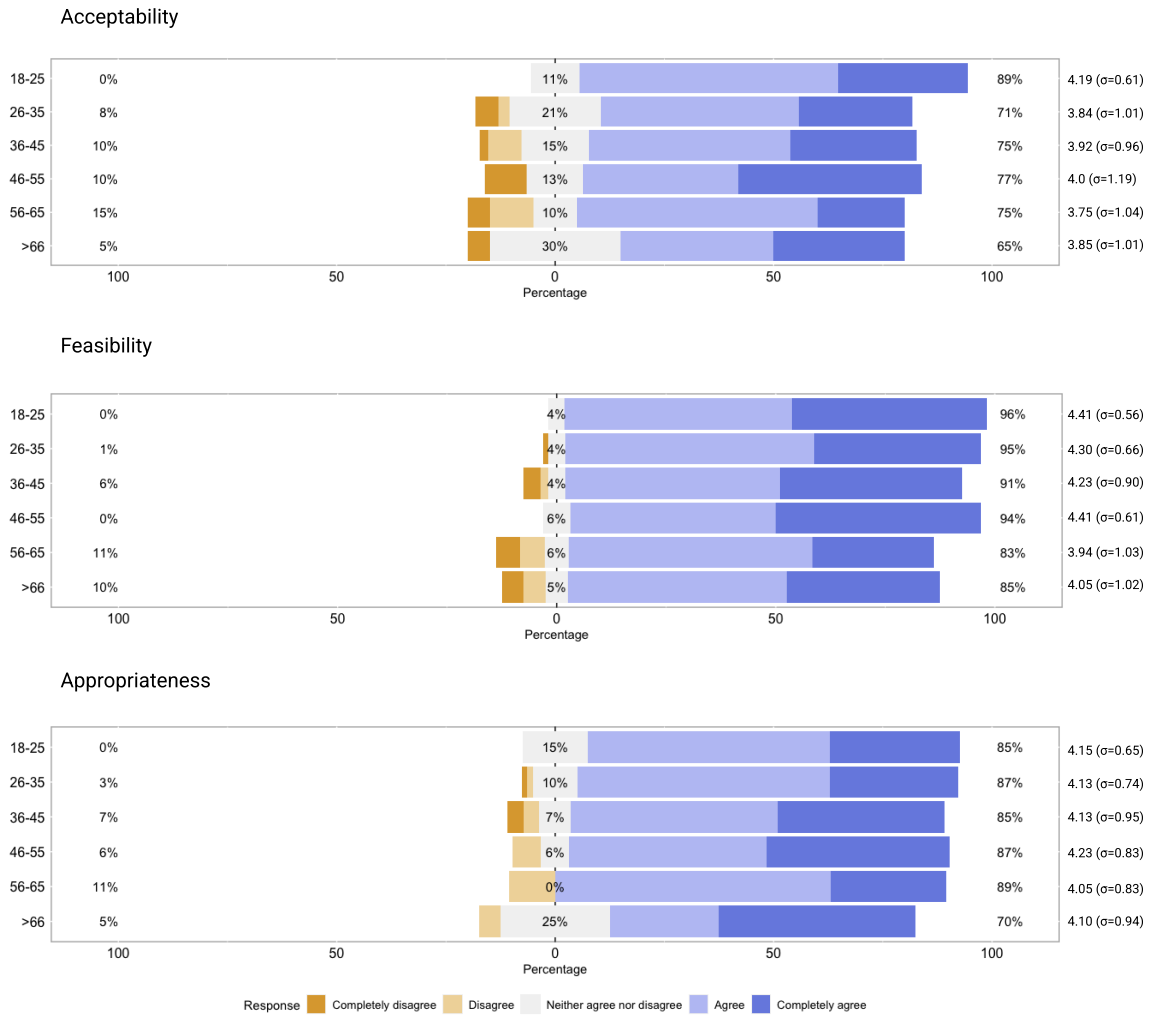


Figure A.1: Diverging stacked bar charts of Likert's scale ratings for acceptability, feasibility, and appropriateness with response distributions by age. The mean and standard deviation for each group are shown on the right.

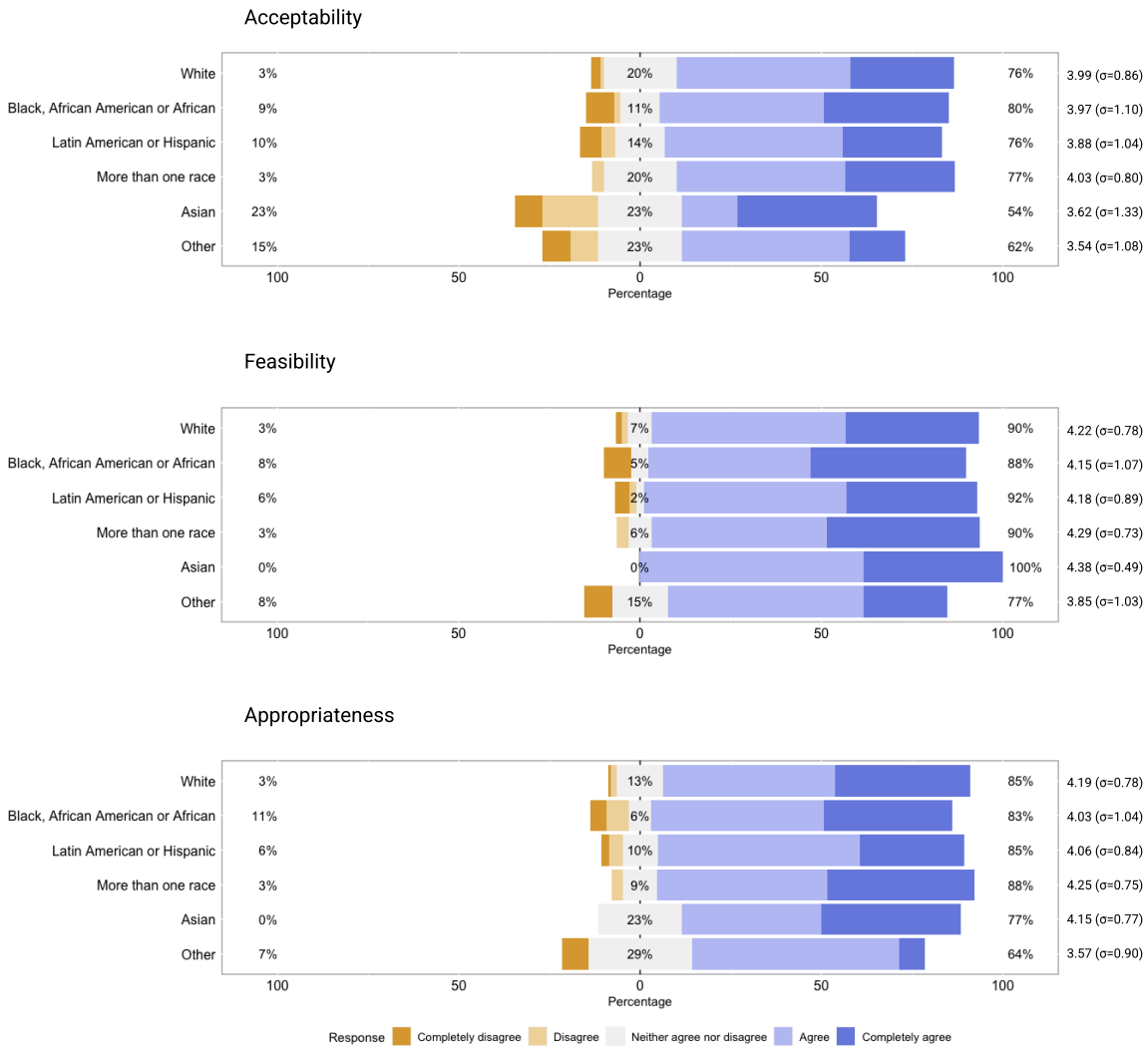


Figure A.2: Diverging stacked bar charts of Likert's scale ratings for acceptability, feasibility, and appropriateness with response distributions by ethnicity. The mean and standard deviation for each group are shown on the right.

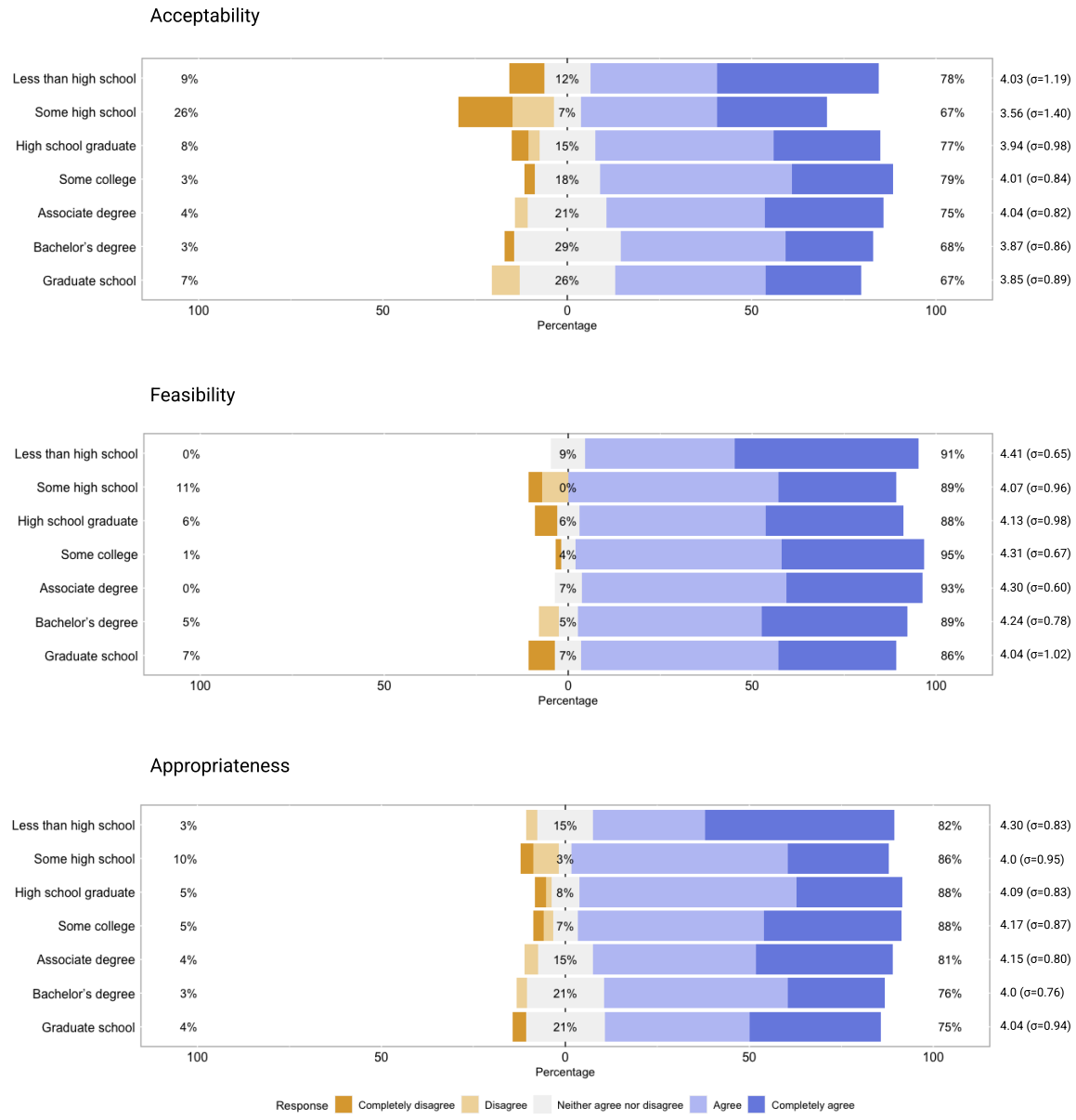


Figure A.3: Diverging stacked bar charts of Likert's scale ratings for acceptability, feasibility, and appropriateness with response distributions by education. The mean and standard deviation for each group are shown on the right.