Heuristic Evaluation of Conversational Agents

Raina Langevin
rlangevi@uw.edu
Human Centered Design and Engineering, University of Washington
Seattle, WA

Ross Lordon
rolordon@microsoft.com
Microsoft
Redmond, WA

Thi Avrahami
thi@rul.ai
Rulai
Mountain View, CA

Benjamin Cowan
benjamin.cowan@ucd.ie
School of Information and Communication Studies, University College Dublin
Dublin, Ireland

Tad Hirsch
tad.hirsch@northeastern.edu
Department of Art + Design, Northeastern University
Boston, MA

Gary Hsieh
garyhs@uw.edu
Human Centered Design and Engineering, University of Washington
Seattle, WA

ABSTRACT

Conversational interfaces have risen in popularity as businesses and users adopt a range of conversational agents, including chatbots and voice assistants. Although guidelines have been proposed, there is not yet an established set of usability heuristics to guide and evaluate conversational agent design. In this paper, we propose a set of heuristics for conversational agents adapted from Nielsen's heuristics and based on expert feedback. We then validate the heuristics through two rounds of evaluations conducted by participants on two conversational agents, one chatbot and one voice-based personal assistant. We find that, when using our heuristics to evaluate both interfaces, evaluators were able to identify more usability issues than when using Nielsen's heuristics. We propose that our heuristics successfully identify issues related to dialogue content, interaction design, help and guidance, human-like characteristics, and data privacy.

CCS CONCEPTS

• **Human-centered computing** → **Heuristic evaluations**; **User interface design**.

KEYWORDS

heuristic evaluation, conversational agents, user interface design

ACM Reference Format:
Raina Langevin, Ross Lordon, Thi Avrahami, Benjamin Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445312

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445312

1 INTRODUCTION

Conversational agents are growing in popularity, through the uptake of text based and voice based conversational systems such as chatbots and Intelligent Personal Assistants (IPAs) respectively. Unlike other forms of human-computer interfaces, there is little consensus as to best practice for the design of conversational agents [5]. Recently there have been strides towards consolidating and validating guidance in related areas, such as human-AI interaction [1], and human-like chatbot experiences [24]. Our work looks to build upon recent efforts [20][26], to develop a comprehensive set of heuristics for conversational agent based interactions. The use of heuristics to guide design and evaluation is a widely used practice for interface design. Our research takes the approach of using Nielsen's heuristics [22] as a foundation upon which to build, adapting these for conversational agent based interaction.

We sought to expand on Nielsen's heuristics using a four phased design process. 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.

In this paper, we contribute 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. 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.

2 RELATED WORK

Conversational agents 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 [8]. Depending on the domain, a conversational agent may be designed for entertainment, companionship, informational or task-based purposes. Conversational agents can also 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.

However, while there is an increased interest in using these technologies, designing conversational agents is not easy. There are a number of barriers to interacting with conversational agents, such as unmatched expectations of the system's capabilities [6], differences in conversation styles [25], increased cognitive load for particular user groups [27] and social embarrassment [7]. Past work has diverged on whether chatbots should exhibit human-like characteristics and a number of desirable human-like behaviors have been proposed [24]. For example, while small talk has been shown to be beneficial for establishing trust [2], it may not be desired based on the context of the chatbot [24] or users' personal preferences [16]. 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.

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 [22], a 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.

However, 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 paper, we focus on validating and adapting Jakob Nielsen's 10 usability heuristics to conversational agents.

2.1 Adapting Nielsen's Heuristics

Heuristic evaluation commonly relies on the set of 10 heuristics established by Jakob Nielsen [22]. 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 [19]. The heuristics were based on their experiences and were supported by the principles outlined in [12] 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 [13] [4] and past work has had success in extending Nielsen's heuristics for smartphones [3], ambient displays [18], and medical devices [28].

There have been recent developments towards heuristics for specific modalities, like voice interactions [26][20]. However, we are not aware of a comprehensive set of heuristics. Due to the lack of validation for design heuristics in specific domains [11], it is important to validate proposed heuristics in line with previous work[1]. In this paper, we utilize a similar design process used in prior work to develop heuristics for ambient displays [18]. We conduct a four phased design process as referenced in Table 1.

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

Table 1: The four phased design process.

3 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).

3.1 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 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 [17]. Therefore, in the last iteration of revisions, we merged our set of heuristics with the adapted set in [17]. 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 [17] we also included Grice's Cooperative Principles [10] 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 [10]. 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 [9].

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 [17]'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 [17]'s set that were targeted towards smartphone modalities. The only heuristic that remained without adaptation was *Help users recognize, diagnose and recover from errors*.

4 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.

4.1 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.

4.2 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.

4.3 Results

As shown in Table 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.

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

5 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.

5.1 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 [15].

5.2 Participants

5.2.1 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.

5.2.2 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.

5.3 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.

5.4 Results

The authors first conducted an informal expert review to generate a master list of all known usability issues, following methodology in past work [18] [22]. 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 and Figure 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

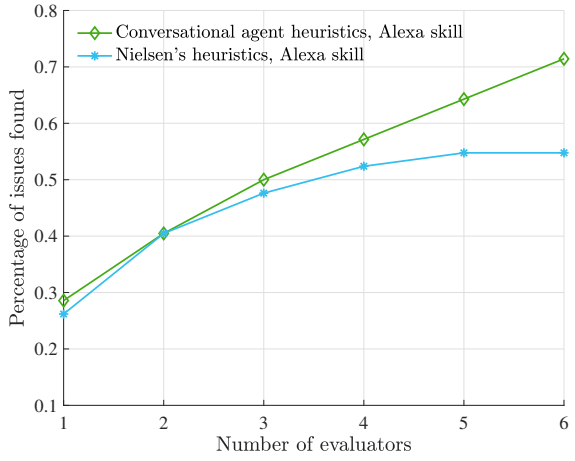
Table 3: Number of usability issues found by the experts, and the top six evaluators in the conversational agent (CA) and Nielsen groups in Phase 3.

5.4.1 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 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.

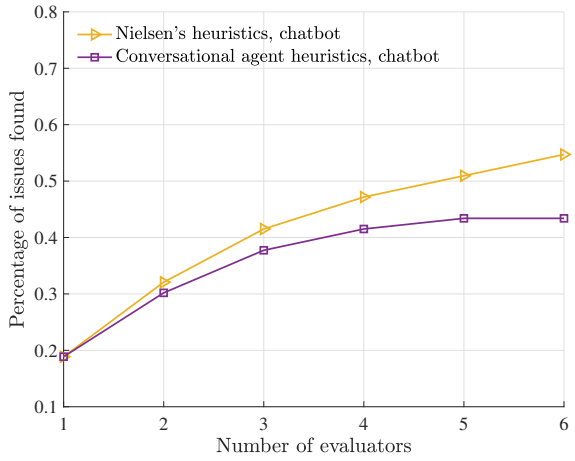
The results showed that the conversational agent heuristics were better able to identify issues than Nielsen's for the Alexa skill. As shown in Table 3, the top 6 evaluators using the conversational agent heuristics identified 30 out of 42 issues compared to those using Nielsen's heuristics 23 out of 42. In Figure 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 [21]. 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.

5.4.2 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



(a) *voice*



(b) *chatbot*

Figure 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.

interface in the online study. In Table 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 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.

5.5 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.

6 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

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 4: The conversational agent heuristics developed in Phase 3.

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.

6.1 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.

6.2 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,

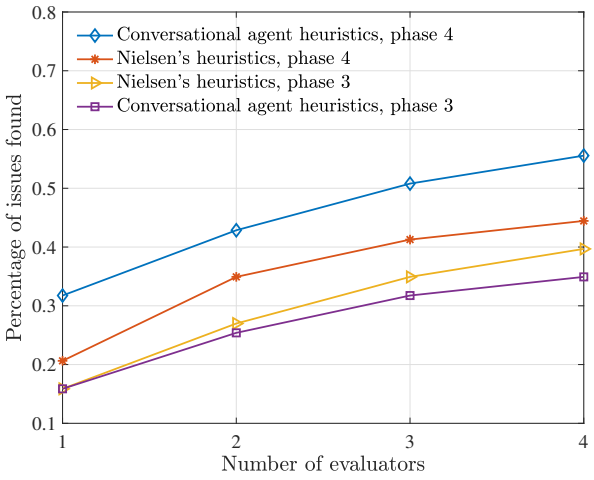


Figure 2: Percentage of issues for the chatbot found by the top four evaluators using the conversational agent heuristics and Nielsen's heuristics in Phase 3 and 4.

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 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 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 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 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 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 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.

6.2.1 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.

6.2.2 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".

6.2.3 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.

6.2.4 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.

6.2.5 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.

6.2.6 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.

6.2.7 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.

6.2.8 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.

6.2.9 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.

6.2.10 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.

7 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. 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 [14]. 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 [23]. 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 [10]. 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.

Prior work has suggested that Nielsen's heuristics are general and do not address relevant areas of specific domains [18][26][20]. 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

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

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*.

While we found that the usability issues identified by the conversational agent heuristics are on average lower than those found by Nielsen, that is mostly because the conversational agent heuristics found more issues, and that these issues are 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

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

heuristics, was given a low rating, but for another type of interface this issue may be more severe.

8 LIMITATIONS

There were additional 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. 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.

9 CONCLUSION

In this work, we proposed and validated a set of 11 heuristics for conversational agents that can be generalized to text, voice and multi-modal 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.

ACKNOWLEDGMENTS

We would like to thank all of the experts who reviewed the heuristics for their invaluable time and feedback. We also thank our study participants who conducted heuristic evaluations and the anonymous reviewers for their helpful feedback. This work was in part supported by the Science Foundation Ireland ADAPT Centre (13/RC/2106).

REFERENCES

[1] 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*. 1–13.

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

[3] Piotr Calak. 2013. *Smartphone evaluation heuristics for older adults.* Ph.D. Dissertation.

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

[5] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The state of speech in hci: Trends, themes and challenges. *arXiv preprint arXiv:1810.06828* (2018).

[6] 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*. 1–12.

[7] 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*. ACM, 43.

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

[9] Google. n.d.. Learn about conversation - Conversation design. https://designguidelines.withgoogle.com/conversation/conversation-design/learn-about-conversation.html.

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

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

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

[13] Keith Instone. 1997. Site Usability Heuristics for the Web. http://instone.org/heuristics.

[14] 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*. 895–906.

[15] 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*, Vol. 2019. American Medical Informatics Association, 552.

[16] 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*. 264–275.

[17] 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. Dissertation. University of Washington.

[18] 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*. ACM, 169–176.

[19] Rolf Molich and Jakob Nielsen. 1990. Improving a human-computer dialogue. *Commun. ACM* 33, 3 (1990), 338–348.

[20] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.

[21] Jakob Nielsen. 1994. How to Conduct a Heuristic Evaluation. https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/.

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

[23] 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 (2019), 785–797.

[24] 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*. 151–161.

[25] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of Style in Information Seeking Conversation with an Agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1171–1180.

[26] Zhuxiaona Wei and James A Landay. 2018. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.

[27] 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* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3405755.3406118

[28] 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 (2003), 23–30.

A APPENDIX

A.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.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.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.4 Phase 2: Expert Review Results

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 9: 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 10: The 12 conversational agent heuristics compared to the earlier modified heuristics, and the average relevance rating for each heuristic.