Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback?

ZIANG XIAO*, University of Illinois at Urbana-Champaign, USA
SARAH MENNICKEN, Spotify, USA
BERND HUBER, Spotify, USA
ADAM SHONKOFF, Spotify, USA
JENNIFER THOM, Spotify, USA

Voice assistants offer users access to an increasing variety of personalized functionalities. Researchers and engineers who build these experiences rely on various signals from users to create the machine learning models powering them. One type of signal is explicit feedback. While collecting explicit user feedback in situ via voice assistants would help improve and inspect the underlying models, from a user perspective it can be disruptive to the overall experience, and the user might not feel compelled to respond. However, careful design can help alleviate the friction in the experience. In this paper, we explore the opportunities and the design space for voice assistant explicit feedback elicitation. First, we present four usage categories of explicit feedback in situ for model evaluation and improvement, derived from interviews with machine learning practitioners. Then, using realistic scenarios generated for each category, we conducted an online study to evaluate multiple voice assistant designs. Our results reveal that when the voice assistant is introduced as a learner or a collaborator, users were more willing to respond to its request for feedback and felt less disruptive. In addition, giving users instructions on how to initiate feedback themselves can reduce the perceived disruptiveness compared to asking users for feedback directly. Based on our findings, we discuss the implications and potential future directions for designing voice assistants to elicit user feedback for personalized voice experiences.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Collaborative and social computing design and evaluation methods**;

Additional Key Words and Phrases: voice assistants, feedback elicitation, machine learning

ACM Reference Format:
Ziang Xiao, Sarah Mennicken, Bernd Huber, Adam Shonkoff, and Jennifer Thom. 2021. Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback?. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 388 (October 2021), 24 pages. https://doi.org/10.1145/3479532

1 INTRODUCTION

In recent years, voice assistants, such as Google Assistant, Amazon Alexa, or Apple Siri, have become increasingly popular. Market research forecasts that 8 billion voice assistants will be in

*The author completed this work as part of an internship at Spotify.

Authors' addresses: Ziang Xiao, zxiao5@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, IL, USA; Sarah Mennicken, sarahm@spotify.com, Spotify, San Francisco, CA, USA; Bernd Huber, bhb@spotify.com, Spotify, Boston, MA, USA; Adam Shonkoff, ashonkoff@spotify.com, Spotify, Boston, MA, USA; Jennifer Thom, jennthom@spotify.com, Spotify, Boston, MA, USA.

2573-0142/2021/10-ART388 $15.00
https://doi.org/10.1145/3479532

use globally by 2023 [1]. Due to the rapid development of machine learning models, voice assistants nowadays deliver more diverse and more personalized content, including daily news briefings, outfit suggestions, and music recommendations. Among voice assistant applications, requesting music is a top use case [7, 10]. Even when the user only provides a simple command, such as *"Play music,"* the services connected to the voice assistant can deliver personalized and context-aware recommendations based on the user's taste and listening history [76].

Researchers and engineers rely on user signals to build quality machine learning models that enable voice assistants to provide personalized experiences [50]. These user signals also help recommender systems to evaluate the recommended content and inform future recommendations. One such signal is explicit user feedback, where users tell the system explicitly about their experience, opinions, or preferences [46]. Compared to other types of user signals, such as interaction logs, explicit user feedback contains less noise [6, 33, 34].

Recent studies found that users are willing to engage with conversational agents and provide meaningful information [87]. This indicates a large, currently under-utilized opportunity to create tighter feedback loops via voice assistants to connect system builders and end users and improve the underlying recommender systems to provide better personalization. However, feedback mechanisms also need to be carefully designed to avoid degrading the user experience [27, 30]. Therefore, without established and well-understood techniques for collecting explicit user feedback, this type of data will not be collected, depriving researchers and engineers of a potential high-quality source of user signal. This missed opportunity might become even more significant as recommender systems — traditionally powered by collaborative filtering — continue their shift towards more responsive algorithms like reinforcement learning, where user feedback plays a central role [29].

Explicit user feedback is especially useful when it is collected in situ [27, 38], i.e., in the place and in the moment in which the interaction of interest occurs. It avoids introducing a memory bias when asking users to recall their impression of an interaction that happened in the past [62]. Also, having access to rich contextual information can help system builders improve recommender systems by tailoring content to specific contexts [4]. For example, users who often listen to slow songs may want something more energetic when going for a run. But collecting explicit feedback in situ is challenging: unlike graphical interfaces where system builders can use pop-up questions or upvote/downvote buttons to prompt for feedback, the lack of visual affordances on voice interfaces limits the system's ability to prompt for user feedback. This "invisible" nature of voice commands [17] makes it difficult for users to even know about the options to provide feedback. Moreover, the voice assistant must interrupt an ongoing interaction to ask for feedback, which can hurt the user experience. For example, if the voice assistant wants to know whether the user likes a song that is playing, the assistant needs to stop the music to ask the user.

Picking the right situations and frequency to ask for feedback is another challenge: frequent feedback requests may cause fatigue that hurts user experience and feedback quality [68]. It is crucial for the voice assistant to ask strategically to reduce experience friction. It should focus on the feedback request that maximizes system gain with minimal user experience friction. To identify what this feedback might be, it is key to involve not only the end users, but also experienced machine learning practitioners. These experts know what user signals would be most beneficial to the development of the underlying systems, and they are well attuned to potential user concerns.

Motivated by the benefits of explicit in situ user feedback and the potential challenges of implementing proactive voice interactions, our research aims to provide an overview and a better

[1] https://voicebot.ai/2019/02/14/juniper-estimates-3-25-billion-voice-assistants-are-in-use-today-google-has-about-30-of-them/

understanding of the potential opportunities for voice assistants to engage with users for feedback elicitation. With our work, we are addressing two research questions:

- **RQ1:** What types of explicit user feedback are valuable to machine learning practitioners?

We conducted interviews with machine learning (ML) practitioners (N = 12) to identify categories of explicit user feedback that are particularly valuable for ML model inspection and improvement. Our analysis revealed four categories, 1) Clarifying user input, 2) Clarifying behavioral signals, 3) Collecting feature feedback, and 4) Understanding user context. By collecting explicit user feedback in situ, a voice assistant could help ML experts evaluate underlying algorithms and enable user models to deliver more accurate recommendations and more engaging experiences. With the understanding of what types of feedback are valuable, we then asked our second research question:

- **RQ2:** How do different voice assistant design dimensions affect user willingness to respond to the feedback elicitation request and the perceived friction to the experience?

We conducted an online study (N = 294) with realistic scenarios to evaluate user perceptions of different voice assistant designs when collecting explicit user feedback. We look at three design dimensions of a voice assistant: 1) Framing: How the voice assistant is being presented to the user — either as an Assistant, a Collaborator, or a Learner; 2) Elicitation Strategy: How the voice assistant elicits the feedback — either as a direct question that the user is expected to respond to, or as an instruction on how the user can initiate providing feedback, 3) Level of Proactivity: When the feedback elicitation occurs in relation to the user's interaction with the assistant — as an extension of an expected response to a voice command, as a response to a user interaction that typically does not involve a response from the voice assistant, or independently from any user interaction.

Our results show that the above design dimensions of the voice assistant are indeed important aspects to consider when aiming to collect explicit user feedback in situ. Users are more willing to respond to voice assistants that are framed as a Collaborator or Learner than to those framed as an Assistant. Voice assistants that provide instructions on how to give feedback are perceived as less disruptive than those asking direct questions. Finally, the choice of the elicitation strategy becomes more consequential as the voice assistant becomes more proactive (i.e., when it attempts to elicit feedback with no direct user interaction preceding it).

Our work contributes in three ways: first, we derived four usage categories through interviews with ML experts in recommender systems and a review of related work. These categories provide potential directions for further investigation to collect explicit user feedback through voice assistants. Second, our online study shows design decisions that can affect a user's perception and experience when a voice assistant is asking for explicit feedback. Lastly, we discuss design considerations and implications for creating voice assistants that can effectively elicit quality feedback in different contexts and mediate collaborations between developers and users in recommender systems.

2 RELATED WORK

2.1 Feedback in Recommender Systems

Signals from the user, such as user behavior logs or explicit user feedback, are crucial in building and evaluating machine learning models. From various user signals, the system can learn and make inferences about a user's interests and preferences [65, 71], and system builders can use those signals to understand when and how the model failed to deliver engaging experiences [77].

Generally, there are two types of user signals: implicit and explicit. Implicit user signals are often derived from user interaction logs, such as the engagement duration, the amount of scrolling on a webpage, skips, or likes. System builders need to make assumptions about implicit signals to understand user intention or preference, which often introduces a lot of noise [33]. In contrast, explicit user signals capture user preferences in a more direct and granular way. Often, such signals

come from user ratings, survey questions, or other mechanisms that allow users to explicitly express their preferences or intentions [49, 65]. Taken together, implicit and explicit signals can provide a comprehensive view of the user experience.

In this study, we focused on one type of explicit user signal: user feedback. Feedback is crucial for developing recommender systems, from early collaborative filtering techniques to deep learning models. These systems learn user preferences through feedback, such as ratings and reviews [71, 91]. Machine learning practitioners also rely on user feedback to inspect system failures [69, 79] or discover user patterns [52].

However, eliciting explicit user feedback is challenging, as it can introduce more effort for the user, and self-reported data is not always reliable [59, 60]. Although people have studied how to elicit user feedback through traditional interfaces [66], using voice assistants to collect feedback spoken by the user to the assistant is underexplored.

2.2 Information Elicitation In Situ

Collecting data from individuals in situ, in the place and in the moment in which the interaction of interest occurs, has proven invaluable in many diverse fields, such as psychology [70], anthropology [5], and human computer interaction [27, 90]. Eliciting user feedback this way can mitigate memory biases during recall [72], ensure the collected data took place within the context of interest [27], and obtain richer contextual information [3].

There are several methods for collecting in situ feedback in research. For example, diary studies in which participants are instructed to note down self-observations that are of interest to the researchers; or the Experience Sampling Method [27, 64], where a system randomly probes and checks if participants are in the context of interest to then collect data in situ. Due to the increasing capabilities of mobile sensors, such systems can sometimes even detect the user context and proactively ask for the feedback. For example, Tallyn et al. developed a chatbot to gather ethnographic data from participants in real-time [74], and Bachmann et al. propose a data collection tool that prompts users with context-aware triggers [8].

Since both implicit and explicit user signals are highly affected by the user's context [9], it is especially valuable to collect user feedback in situ. To do so, front-end developers and user interface designers often include feedback buttons, such as Like buttons, or text input boxes for suggestions on the interface [46]. However, such passive methods are often ignored by the user or result in data biased towards people with complaints [6]. To make the feedback collection more noticeable, sometimes these prompts for user feedback are triggered by certain user interactions [8, 74]. However, such proactive methods can create a significant amount of unwanted friction to the user experience or unnecessary user burdens.

In the context of voice assistants, it is nearly impossible to passively collect user feedback, as we cannot just add a visual feedback button to a voice interface. Even if the voice assistant is on a device with a screen, we cannot be sure the user's focus of attention is on the screen without additional forms of measurement (e.g., eyetracking) [51]. But given the value of explicit in situ user feedback, we set out to explore the potential design space for proactive information elicitation via voice assistants. For this, we evaluated different designs to identify design decisions that can alleviate some friction and create more willingness in the user to respond to the assistant.

2.3 Conversational Interfaces for Information Elicitation

Conversational user interfaces allow users to interact with computer systems in natural language, providing a more flexible [78] and personalized user experience [32]. The prospect of having access to more natural user input encourages researchers and practitioners to explore the use of conversational interfaces for information elicitation [87].

Compared to traditional, static online surveys, conversational interfaces can enhance information elicitation by providing interactive feedback and asking follow-up questions [18]. Xiao et al. showed that compared to a form-based survey, their interview chatbot elicited more relevant and informative responses, while creating a more engaging user experience [87]. Research also found that people are willing to disclose information about themselves to conversational agents [47]. Lucas et al. found that the participants who believed they were interviewed by a computer were more willing to disclose information than those who believed they were interviewed by a human interviewer [47]. Promising results from prior studies encourage us to further explore the use of voice assistants, another popular conversational interface, to collect explicit user feedback.

However, prior research also shows that users have privacy concerns when voice assistants elicit or handle sensitive data [1, 15, 19, 23]. For example, Cho found that soliciting sensitive health information through a voice assistant raises privacy concerns which negatively impacts users' general perception of the assistant [15]. Such concerns increase further when the social context is inappropriate. Moorthy and Vu found that users felt more comfortable interacting with voice assistants (e.g., Siri) in private settings than in public spaces where others at the scene may overhear the user's conversation [23]. Privacy concerns not only reduce the usage frequency but also deter users from using certain functions altogether [1, 19]. Given this mix of potential benefits and privacy threats in information elicitation via conversational interfaces, it is crucial to consider both the value of the user's feedback to the developer and the cost to the user experience when assessing the user perception of the voice assistant's feedback request.

Our work expands previous studies in two ways: first, unlike prior work that focused predominantly on text-based chatbots or embodied virtual agents, our work focuses on voice assistants, an increasingly popular conversational interface. Second, the voice assistant designs in our study collect user feedback in situ, integrated as part of the interaction as opposed to eliciting information after the fact.

2.4 Designing the Social Aspects of a Voice Assistant

From a technical implementation standpoint, voice assistant interactions consist of voice input from a person and synthesized speech output from the device. However, the Computers are Social Actors (CASA) paradigm suggests that people's relationships with computers are fundamentally social, which means humans engage with computers in a manner similar to how they engage with each other [58]. Nowadays, voice assistants have been deeply embedded in our daily life, and the interaction design of a voice assistant should be accountable and appropriate for various social situations [67]. Moreover, designers need to make numerous decisions about the assistant's voice and behavior to design a voice assistant that will resonate with users [13]. These decisions are especially crucial when designing an experience where users and a voice assistant engage in social interactions, such as prompting and providing feedback.

One such decision is how to frame or introduce the voice assistant to the user. Designers can frame a voice assistant using metaphors with social characteristics. For instance, framing a conversational assistant by presenting dimensions of warmth and competence can influence people's willingness to interact and cooperate with agents, even if they signal low competence [36]. Braun et. al also explore in-car voice assistant designs on the dimensions of Dominant – Submissive and Hostile – Friendly and suggests if the voice assistant character matches the user's personality it will encourage trust and engagement [11]. The right framing for the voice assistant is important since social features for non-living actors can deceive users into overestimating the assistant's capabilities, causing frustration when encountering limitations in actual use [53, 54]. Finally, the impact of framing extends beyond voice assistants, as studies on other forms of conversational agents also suggest effects of the agent's framing on the user experience [86].

Another set of design decisions involves how the assistant will act and react to users through specific conversational and elicitation strategies. Users will explore the limits of their voice assistants by issuing playful commands [48]. In autonomous vehicles, small talk to build rapport can foster trust between the voice assistant and the user [43]. Voice assistants can also shape work practices by highlighting social information elicited from users (e.g., where people deviated from their routine) [14]. Whether the interaction is led by the assistant or the user can also affect user perceptions of the voice assistant [56].

Most current interactions with voice assistants are rather transactional, where the user issues a request and the assistant provides a response [7]. However, more recent work has started to explore opportunities for voice assistants to play a more proactive role, e.g., by trying to predict opportune moments to initiate voice interactions when driving [37], or by empowering users to have meaningful interactions with their own usage data [84].

In our work, we address these different approaches and design choices, extending research by focusing on voice assistant framing (e.g., Assistant, Collaborator, or Learner), options for different elicitation strategies, and the effects that different proactivity levels have on the perception of feedback elicitation via voice.

3 METHOD OVERVIEW

Our work aims to address two research questions. In RQ1, we identify different types of explicit user feedback that are valuable to ML practitioners. We then use these insights to investigate RQ2, learning how voice assistant design decisions can affect users' willingness to respond to the feedback elicitation and the perceived friction to the experience.

To address RQ1, and to lay the foundation for a study that would allow us to explore RQ2, our first step was to conduct expert interviews with ML practitioners. Our second step, to address RQ2, was to run an online study in which we applied different voice assistant design decisions to realistic scenarios that were generated based on insights from the interviews and related work.

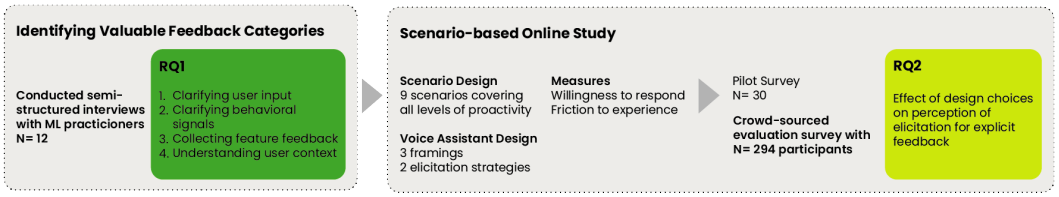

Fig. 1. Overview of the two steps of our research: from identifying valuable feedback categories through expert interviews to generating scenarios evaluating the effect of different design choices on the user perception.

4 IDENTIFYING VALUABLE FEEDBACK CATEGORIES THROUGH EXPERT INTERVIEWS

To identify categories of explicit user feedback that are valuable to collect via voice assistants, we conducted semi-structured in-depth interviews with machine learning experts who design, develop, or evaluate machine learning models that directly shape the interactions of end users. We recruited 12 machine learning practitioners (six women and six men) from a content-streaming company for which recommender systems play a central role in most of its products. Six participants have directly worked or are working on recommender systems or user models that interact with users through voice assistants. We conducted semi-structured interviews of approximately 45 minutes with each

participant via a video conferencing software [2]. At the beginning of each interview, we asked the participants about their roles, past experience, and challenges in building and evaluating machine learning models. Then we focused on their opinions and thoughts on explicit user feedback in situ in the context of their work. For example, we asked what kinds of user feedback are particularly useful to them and, if applicable to their experience, we asked specifically about collecting feedback through conversational voice interactions. All interviews were transcribed and then analyzed by two authors through inductive coding and clustering to identify common categories of scenarios in which such feedback would be useful to machine learning practitioners.

4.1 Usage Categories

Through our interviews, we identified four usage categories that represent a broad range of scenarios in which explicit in situ user feedback through conversational voice interactions could benefit the underlying machine learning models or recommender systems: (1) Clarifying user input, (2) Clarifying behavioral signals, (3) Collecting feature feedback, and (4) Understanding user context. In the following, we will briefly describe these categories.

4.1.1 Clarifying user input. This category describes scenarios where the user issues a query to the system but the query is too ambiguous for the system to make a decision with high confidence. The voice assistant can then prompt the user for clarification. One machine learning expert (P2) gave the following example,

> *"If somebody says ... Excuse my vulgarity, but if somebody says, 'F**k you.' to a voice assistant, are they upset with the voice assistant, or are they looking for a song by CeeLo Green? It's really tough to just look at just those two words, and be able to say, 'I am 100% [sure, I] know what that is.' "*

In this example, a voice assistant could disambiguate the user input by proactively prompting the user to clarify their request, saying, "Do you want the song by CeeLo Green?" The interviewed machine learning practitioners described this usage category as particularly valuable, as "[user] utterances are short and ambiguous" (P2). The relevance of this category is further backed by prior studies, which found that clarifying ambiguous user input can help deliver content that best matches the user's needs [35, 49, 63]. Another benefit mentioned by the machine learning practitioners is that such explicit user input in response to the voice assistant's clarifying question can help create higher quality datasets with less noise. For example, participant (P4) said,

> *"Most of the ambiguity could be cleared up I think, if we could just ask users questions. That would really help us understand which rows have labels that are pure noise and which rows have labels that are reliable."*

4.1.2 Clarifying behavioral signals. The second common theme mentioned is about clarifying behavioral signals. This refers to occasions when a user's behavioral signals — such as duration of user engagement, skipping content, or completing content to the end—might contradict each other. One participant (P5) mentioned,

> *"Follow[ing][3] and play[ing an artist or podcast] could be contradictory. It seems like people tend to follow something, which one might think means they are really interested in it, but [then] they don't actually engage with."*

Nowadays, machine learning systems, especially reinforcement learning systems, heavily rely on users' behavioral signals to reward or punish the system actions and eventually learn user

[2] Please refer to the supplementary material for the interview protocol.

[3] *Following* in this context means subscribing to notifications about an artist's or a podcast show's new releases.

preferences. Here, the voice assistant could proactively ask for explicit feedback when detecting such ambiguity in the behavioral signals. Clarifying contradictory signals could not only help the models to infer user preference with confidence, but it could also help machine learning experts to discover new user patterns [88].

4.1.3 Collecting feature feedback. When introducing new features, it is often initially unclear what implicit signals are most meaningful or reliable. For example, one machine learning expert (P2) spoke about having to come up with evaluation metrics before knowing how users will respond to it,

> *" ... Because you can't really collect any data from users if there's no feature to actually collect that data from. So that was the initial challenge that we faced."*

In cases like these, user researchers are often brought on to conduct user studies. However, our participants and previous studies mentioned that this kind of user research often takes a long time, and the data they provide is retrospective and disconnected from the data used to build the model. Voice assistants might open new opportunities to prompt the targeted collection of feedback on early feature prototypes at the moment of interaction and with a faster turn-around time. Collected feedback could be immediately available to the machine learning experts and directly linked to the relevant interaction logs or model output facilitating the creation of reliable evaluation metrics.

4.1.4 Understanding user context. Lastly, our participants reported that by taking the user's context into account, many user models and recommender systems could be improved. For example, in the music context, the activity a user is currently engaged in, or whether they are with other people at the moment, can strongly affect what they might want to listen to [24]. One expert (P12) states,

> *"I think [collecting user context] is pretty challenging because where our user is right now, whether people are with them or not, is really difficult to figure out, but at the same time, very important."*

Our participants also raised privacy concerns on collecting user context, as expressed by P6:

> *"I think a lot of the hurdles that we face on that front are more around privacy than anything else, which I greatly respect. "*

Thus, the system should acquire a user's consent prior to collecting such data. We also learned from our interviews that some contextual data, such as mood or the social setting of a user, can be technologically difficult to capture and thus relies on users' self-reporting. Such challenges were also mentioned in previous literature [4, 27]. Eliciting explicit feedback from the user could help alleviate both the privacy as well as technological challenges. We believe that if voice assistants directly ask the user for such information, it might improve the transparency about what data is gathered, implicitly get the user's consent, and facilitate context detection that might be technologically challenging or inherently subjective.

5 SCENARIO-BASED ONLINE STUDY

Our online study directly addresses RQ2, making use of scenarios we generated for the identified usage categories to understand how certain design decisions affect users' willingness to respond and the perceived friction to the experience. We chose an online scenario-based survey for several reasons: first, an online survey allows us to reach a large and diverse participant pool. Second, with a large group of participants, we can compare many design outcomes in multiple scenarios with sufficient statistical power. Third, presenting the scenarios to participants as text-only protects us from unwanted confounding individual effects of the assistant voice itself, which is known to be a major factor [61]. Lastly, online scenario-based studies have been successfully used to investigate user perceptions in other contexts [20, 28].

Category	Example Scenario	Voice Assistant Prompt Variations
Clarifying user input	You say to your assistant: "Hey VA, play Thriller."	[Question] Did you want the song Thriller, or the album Thriller? [Instruction] Okay, here's the song Thriller by Michael Jackson. If you wanted the whole album, say, "play the album Thriller."
Clarifying behavioral signals	You say to your voice assistant: "Hey VA, play me some Beatles music." The voice assistant plays "Hey Jude" by the Beatles, which you've listened to many times in the past. You really like the song, but this time you want to try something else. So you say: "Hey VA, skip this song."	[Question] Oh okay, did I not play what you wanted? [Instruction] Sure, here's another Beatles song. If you want to tell me why you skipped 'Hey Jude', say something like, "Hey VA, I skipped because I wanted to hear something new."
Collecting feature feedback	You feel like hearing some music you've never heard before. You say to your voice assistant: "Hey VA, recommend some music." Your voice assistant says: "Okay, check out this new release by Arianna Grande, an artist who you have listened to a lot in the past." Then it plays that song.	[Question] After the song ends, the voice assistant says, "What did you think? Was hearing my explanation before the song useful?" [Instruction] After the song ends, the voice assistant says, "If you want to tell me if my explanation before the song was useful, say 'Hey VA, I have some feedback.'"
Understanding user context	You're feeling like you want some music but not sure exactly what to play. You say to your voice assistant: "Hey VA, play some music I'd like."	[Question] Sure, I can give you better suggestions if you let me know your mood. How are you feeling at the moment? [Instruction]Sure, I can find something based on your mood if you want. You can say things like 'I'm happy' or 'I'm feeling kind of down.'

Table 1. Overview of the four usage categories with an example scenario for each and two voice assistant prompt variations.

5.1 Scenario Design

For each of the scenario categories, we generated scenarios based on examples mentioned by our participants in the interview study (see Tab 1). Overall, we included 9 scenarios: 3 for Clarifying user input, 2 for Clarifying behavioral signals, 2 for Collecting feature feedback, and 2 for Understanding user context. Each of the scenario descriptions consists of three parts: setting the interaction context, the interaction itself, and a prompt by the voice assistant.

While our scenario generation process made sure that the scenarios were realistic from a machine learning engineer perspective, we wanted to ensure the same from the user perspective. Thus, inspired by Gabriele et al. [28], we introduced a validity check in our study, asking participants to rate how plausible and how likely to happen the described scenario is, each on a 3-point Likert

Framing	Voice Assistant Introduction
Assistant	Imagine you have a voice assistant that helps you with your daily tasks.
Learner	Imagine you have a voice assistant that's always learning how to better help you with your daily tasks.
Collaborator	Imagine you have a voice assistant that collaboratively works with you on your daily tasks.

Table 2. Introductions for the three different voice assistant framings in our study.

Scale (1-3). Our participants considered all scenarios to be plausible (M = 2.47, SD = 0.73) and likely to happen in their daily interactions with a voice assistant (M = 2.46, SD = 0.73), without any significant difference among the four categories. This suggests that users perceived our scenarios to be realistic.

5.2 Voice Assistant Design

Each participant was assigned to one voice assistant design condition. In this section, we describe the design decisions we wanted to investigate in our study and how we modified our scenarios to do so.

5.2.1 Framing. Previous research has shown that people attribute human-like interpersonal behaviors to AI agents [58]. There is evidence that in human-AI interaction, the personality or the metaphor of an intelligent agent affects how people perceive and interact with the agent [11, 36]. Thus, we wanted to investigate the role of a voice assistant's framing on users' perception of user feedback elicitation. For that, we introduced the voice assistant in three different framings: (1) as an Assistant, (2) as a Learner, and (3) as a Collaborator.

The Assistant condition represents how voice assistants are typically presented – framed as virtual helpers. The Learner condition was designed to signal the limited initial capabilities of a voice assistant, as prior studies showed that people are more friendly to a virtual agent when they know the virtual agent is imperfect [36]. Lastly, in the Collaborator condition, our framing echos prior studies where framing human-AI interactions as collaborative tasks or creating shared goals would induce positive attitudes toward the virtual agent and lead to better outcomes [57]. Before reading the scenarios, each participant was primed with an introductory sentence that clearly highlighted one framing as shown in Table 2.

5.2.2 Elicitation Strategy. Elicitation Strategy is a key component when collecting high-quality explicit user feedback in situ. Different feedback prompts may induce different levels of friction, which ultimately affects the quality of user feedback. Multiple studies have indicated that virtual assistants should be adaptive in terms of how they pose questions or requests [86, 89]. In our study, we considered two elicitation strategies: (1) directly asking the question, and (2) providing instructions for the user to provide feedback to the voice assistant. We created two variations for each interaction scenario. For example, when getting ambiguous user requests like "Play Thriller", the voice assistant could ask directly:

> *"Did you want the song Thriller or the album Thriller?"*

or it can provide instructions

> *"Okay, here's the song Thriller by Michael Jackson. If you wanted the whole album, say 'play the album Thriller.' "*

5.2.3 Level of Proactivity. Each scenario in our study is further characterized by its Level of Proactivity. While in all our scenarios the voice assistant initiates the prompt to elicit the feedback, in some this prompt is directly connected to a user input that preceded it, while others are more removed from a specific user interaction. In many scenarios, this is less of a design choice, but a result of the nature of the interaction.

To understand the effects of these different circumstances, we defined three levels of proactivity: (1) A low level of proactivity describes scenarios where the voice assistant's question directly follows a user's command and asks for information that is closely related to the current session. For example, the voice assistant is asking for clarification when the user's initial utterance is ambiguous, as in the "Thriller" example described above. (2) A middle level of proactivity describes scenarios where the voice assistant's prompt is not a direct response to a user's request to the assistant, for example when the user skips multiple songs or "likes" a song. Finally, (3) a high level of proactivity describes scenarios where the voice assistant initiates a new session with a prompt that is not triggered by the user's interaction, for example asking the user if they liked a song once it has finished playing.

Based on the voice assistant's Framing and Elicitation Strategy, we yielded a 3×2 experiment design with six design variations.

5.3 Measures

To understand the participants' perception of the voice assistant's prompt for explicit feedback, we measured two major aspects: willingness to respond and friction to the user experience.

5.3.1 Willingness. We assessed participants' willingness to respond to the voice assistant's prompt for explicit feedback. Studies have shown that the strength of a person's willingness to act indicates the likelihood of actual behavior, such as adopting new technologies [75] and providing feedback [31]. In our case, we want to know how people may respond to the feedback request from a voice assistant with different designs. Based on previous studies that measured people's behavioral intention [73], we also used a 4-point Likert scale and asked a similar question, "How would you react to what the voice assistant said?" The 4-point scale ranged from "It's extremely likely that I would say nothing."(1) to "It's extremely likely that I would say something back to the assistant."(4) The question directly measures participants' willingness to respond to the voice assistant's prompt. As recommended by [22], the item is clearly worded, free of jargon and easy to understand. To ensure that the study participant properly read and understand the scenario and the voice assistant's prompt, we asked a follow-up question, "If you were to say something back to the voice assistant, what would you say?"

5.3.2 Friction. We then evaluated the friction that such voice assistant prompts would introduce to a user's experience. The user perceived experience friction indicates the cost of a voice assistant's feedback request. Assistant design that can induce lower friction may collect better feedback quality and achieve better user retention [87]. Also, previous work suggests a potential link between a user's perceived friction and the assistant's level of proactivity [85], which we also considered in our voice assistant's design. We adapted three 5-point scale questions used by prior studies to the context of our scenarios. Those questions have been shown to be effective indicators of friction [83]. First, we asked directly about the perceived experience friction: "How disruptive to your overall experience is the voice assistant speaking at that moment?" Second, we assessed the friction indirectly by asking for the desired frequency of such interactions: "In a scenario like this one, how often do you think a voice assistant should respond the way it did?" And third, we assessed the perceived cognitive load: "How hard was it for you to figure out what to say in your response to

the voice assistant?" All three questions are 5-point Likert Scale questions, ranging from 1 (Not disruptive at all/ Never/ Extremely easy) to 5 (Extremely disruptive/ Always/ Extremely difficult).

5.3.3 Demographic Measures and Individual Characteristics. As prior literature suggests that certain demographic information and pre-existing attitude correlate with people's perception of technology [55, 82] and their experience with voice assistants specifically [11], we also collected basic demographic information, participants' personalities, and their attitude to and experience with voice assistants.

The basic demographic information included age, gender, education level, and annual household income. Participants' personalities were measured by TIPI, a well-examined 10-item personality scale [25] where participants rate ten sentences about their personality on 7-point Likert scales.

To assess the participants' prior voice assistant experience, we asked them what voice assistants and devices they use or have used in the past. They further self-reported their interaction frequency with their voice assistant(s) on a scale from "Multiple times per day" to "Never". Based on the Technology Acceptance Model [80], we collected participants' existing attitude towards voice assistants from the perspective of Usefulness and Satisfaction on 5-point Likert scales (from -2 to 2). To also assess their trust towards voice assistants, we added a single-item scale, similar to [11].

5.4 Study Procedure and Participant Recruiting

Our study had four sections: In the first section, we asked our participants about their prior voice assistant experience and their attitude towards voice assistants (see Section 5.3.3). For the second section, each participant was randomly assigned to one of the six potential combinations of three different framings (Assistant, Learner, or Collaborator) and two elicitation strategies (asking questions or providing instructions for giving feedback). Participants were asked to read the Framing, a short paragraph of text as described earlier in Section 2. To ensure that participants read the passage, we asked them to rephrase how they would describe a voice assistant and used it as an attention checker. In the third section, we presented each participant with four randomly selected scenarios from the generated scenario pool (see Table 1), one from each of the four usage categories. For each participant, the voice assistant followed the same elicitation strategy in all scenarios. After each scenario, we asked the participants to answer a short questionnaire about their reaction and perception of the voice assistant's prompt. In the last section, the participants were asked to answer a series of demographic questions.

In the survey, participants read a total of four scenarios, one after another and for each scenario, they answered a short series of eight questions (see example question in Figure 2).

Before deploying our survey to a broader audience, we ran a pilot study with 30 participants recruited from Appen with the same criteria as the actual study (described below). In the pilot study, we explicitly asked our participants about the clarity and validity of the generated scenarios, and about confusion around the experiment procedure and question wording. We then revised our study based on participants' feedback.

For the actual study, we recruited our participants from Appen [4]. This platform has been widely used to gather human intelligence in AI research and social science experiments [21, 81]. We sent out our survey in five batches over the course of a week: three during weekdays and two on the weekend to recruit a larger variety of participants. Participants were paid $12.5/hr regardless of their performance and spent approximately 20 minutes on the survey (M = 19.03 mins, SD = 18.31 mins). Since our study materials were written in English, our study targeted participants only from

[4] https://appen.com/

	Question	Instruction	Total
Assistant Condition	48	48	96
Learner Condition	57	50	107
Collaborator Condition	45	46	91
Total	150	144	294

Table 3. Breakdown of the conditions our participants were assigned to based on a combination of Framing and Elicitation Strategy.

English-speaking countries [5]. To ensure quality, we made our study limited to people who are qualified as high-quality workers based on their history with Appen [6].

5.5 Limitations

The goal of this study is to explore the design space for eliciting explicit user feedback using voice assistants. Given the inevitable cost to the user experience when prompting for feedback, we wanted to start with scenarios that would bring the most benefits to machine learning practitioners. Using a scenario-based study design allowed us to evaluate multiple design dimensions effectively and gather input on which ones to pursue further, e.g., as prototypes. However, the hypothetical nature of the scenario-based design also introduces some limitations that should be addressed in future work. First, although our validity check showed that all scenarios in the study are perceived to be plausible and likely to happen for the participants, it still requires them to imagine being in that situation. Second, we measured users' self-reported willingness to react and the perceived disruptiveness to the experience instead of their actual behavior. Thus, as a next step, we would develop prototypes to evaluate the scenarios as tangible experiences in the field and measure users' more natural behavior directly.

The generalizability of our study is also limited by the scenario design focusing on one particular application domain, i.e., music. Voice assistant users hold different privacy norms across usage domains [2]. This means that in other application domains of recommender systems, such as health [1] or finance [45], people might consider a voice assistant's feedback request more as a threat to their privacy. Although one of our study scenarios asked for sensitive data such as the user's context and their emotional state [39], it will require further investigation to understand how people perceive and react when the voice assistant is asking for more sensitive information.

Lastly, we want to acknowledge the absence of text-to-speech voice in our study design. The different characteristics of text-to-speech voices can have a strong impact on people's preferences and behavior [40], which may influence how a user reacts to a request from a voice assistant. However, in a large-scale online study, it is challenging to make our voice assistant's voice adaptive to every participant's preference. Any mismatch in our voice choice with the participant's preference may create unwanted effects that undermine our results. Therefore, in the future, it is important to understand people's reactions to voice assistants' feedback requests in a real-world setting where the participant is using their own voice assistant. Additionally, it is important to understand how different text-to-speech voice choices may affect an assistant's feedback elicitation and the interplay with the design dimensions examined in our study.

[5] https://en.wikipedia.org/wiki/List_of_territorial_entities_where_English_is_an_official_language
[6] https://success.appen.com/hc/en-us/articles/203219195-Guide-To-Contributors-Channels-Page

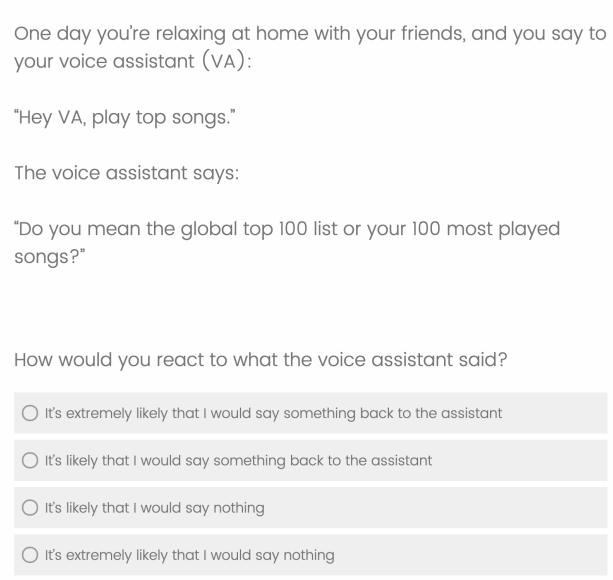

Fig. 2. Screenshot of an example scenario and one of the survey questions as shown to participants.

6 RESULTS

Our results suggest that the choice of Framing and Elicitation Strategy impacts the user's willingness to respond and the perceived disruptiveness, while the Level of Proactivity moderates these effects. In this section, we will first provide an overview of the data we collected, and then present our findings in detail.

6.1 Analysis

We analyzed our data using linear mixed effects method for its robustness in modeling non-independent data. Since each participant in our study responded to four scenarios with the same set of questions, we treated each participant in the linear mixed effects model as a random effect [26], and thus each participant's response to each scenario is a single data point in our analysis. This results in 1,176 (294 participants $\times$ 4 scenarios per participant) data points in our dataset.

We implemented Satterthwaite's method in our linear mixed effects model to estimate the effective degrees of freedom [26]. We tested assumptions including linearity, homogeneity of variance, and the normality of residuals. All p-values reported in the results are adjusted via the Bonferroni correction.

We treated each measure as a dependent variable. For each scenario, the independent variables were the Framing of the voice assistant, its Elicitation Strategy, and the scenario's Level of Proactivity. We also included participant's demographic, personality, and prior attitude towards voice assistants as covariates to control for potential confounding effects suggested by prior literature [11, 42, 82].

6.2 Participant Overview

Out of the 356 participants we recruited, 294 completed the study and passed our attention check question. Our analysis is based on those 294 valid responses. Participants were randomly assigned to one of the six conditions (see Table 3). Among those 294 participants, 126 identified as women,

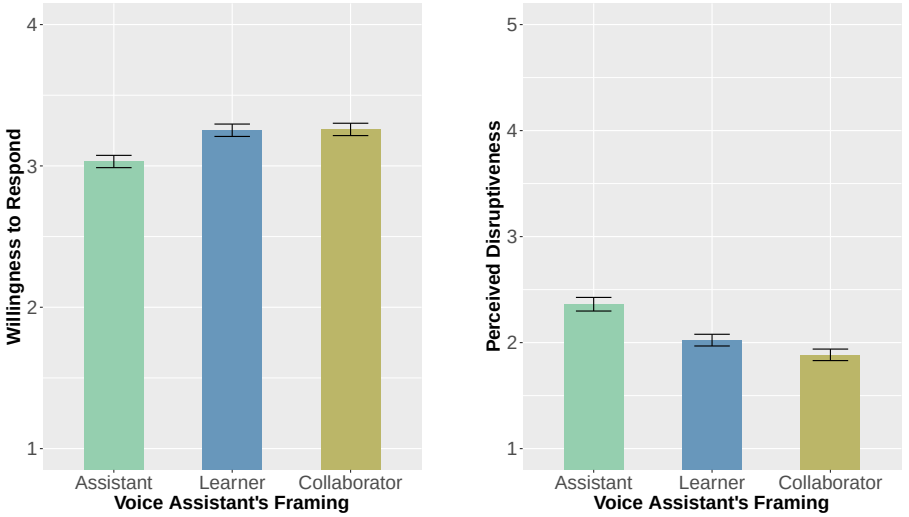

Fig. 3. Differences in participants' willingness to respond (4-point Likert scale) and perceived disruptiveness (5-point Likert scale) for different voice assistant Framings. Overall, people were more willing to respond to a voice assistant's feedback request when it is framed as a learner or a collaborator. And people perceived a voice assistant's feedback request as less disruptive when it is framed as a learner or a collaborator.

145 identified as men, and 23 preferred not to disclose. The median household income was $25,000 - $50,000. The median education level was a Bachelor's degree. The participants' ages ranged from 18-60 years old (M = 28.01, SD = 6.61).

All 294 participants currently owned or used to own at least one voice assistant (e.g., Google Home/Google Assistant, Amazon's Echo/Alexa, Apple's HomePod/Siri, Microsoft's Cortana, Samsung's Bixby). 62.59 % of the participants indicated that they interact with voice assistants at least once per day. 5.44 % of the participants said they had no recent interaction with a voice assistant. The results from the adapted Technology Acceptance Model questions showed that our participants consider voice assistants generally to be Useful (M=1.10, SD=0.74), Satisfying (M=0.60, SD=0.47), and Trustworthy (M=1.00, SD=1.01).

6.3 Higher willingness to respond to a learning or collaborating voice assistant

We found a significant main effect of the voice assistant's Framing: participants were more willing to respond when the voice assistant was introduced as a Learner (M = 3.25 , SD = 0.91), compared to when the voice assistant was introduced as an Assistant (M = 3.03 , SD = 0.85; β = 0.23, SE = 0.09, t = 2.50 , p < 0.05*). Meanwhile, participants are also more willing to respond when the voice assistant was introduced as a Collaborator (M = 3.26 , SD = 0.83), compared to when the voice assistant was introduced as an Assistant (M = 3.03 , SD = 0.85; β = 0.28, SE = 0.09, t = 2.90 , p < 0.01**). Our model showed no significant difference between the Collaborator framing and the Learner framing. We found no significant effect of the Elicitation Strategy and the Level of Proactivity on the willingness to respond.

We also looked at the effect size of the Framing on participant's willingness to respond. We found a small effect with a Cohen's d of 0.38 between the Learner and Assistant framing, and a Cohen's d of 0.39 between Collaborator and Assistant. Figure 3 shows the differences in participants' willingness to respond between the different conditions.

6.4 Less perceived disruption by a collaborating or learning assistant

We found a significant main effect of the voice assistant's Framing on the perceived disruptiveness. Participants perceived the voice assistant to be less disruptive when it was framed as a Learner (M = 2.02 , SD = 1.14), compared to when the voice assistant was framed as an Assistant (M = 2.36 , SD = 1.26; β = -0.39, SE = 0.11, t = -3.56 , p < 0.01***). Furthermore, participants perceived the voice assistant to be even less disruptive when it was framed as a Collaborator (M = 1.89 , SD = 1.04; β = -0.44, SE = 0.12, t = -3.76 , p < 0.01***), compared to the Assistant condition (M = 2.36 , SD = 1.26). However, we found no significant difference between the Learner condition and the Collaborator condition.

We found a small effect with a Cohen's d of 0.39 between Learner and Assistant framing, and a medium effect with a Cohen's d of 0.57 between Collaborator and Assistant. Figure 3 shows the differences in perceived disruptiveness between the different conditions. There were no significant effects of the voice assistant's Framing on the other two measures of friction, acceptable frequency and perceived cognitive load to answer.

The results are consistent with prior research on the CASA paradigm [58], where people responded to computer systems as though the computers were social entities. When the voice assistant is framed as a Learner, it signals initial limited capabilities and the agency for improvement. Similar to what has been shown in [36], signaling the incompetence of artificial intelligence makes people more tolerant to its shortcomings. In our case, the voice assistant signaling incompetence by being introduced as a Learner, might make users feel like teachers, and as such feel more inclined to provide feedback to their "student"—the voice assistant. Moreover, signaling the agency to improve may make the user feel that their feedback will ultimately enhance their experience, which increases the voice assistant's chance to collect feedback without disrupting much.

The Collaborator framing emphasizes a shared goal of creating a better user experience for the user. Such a shared goal can make users more cooperative with an agent's requests [57]. This, in turn, might have led to the observed higher willingness to respond.

6.5 Instructions are perceived as less disruptive than questions, without a negative impact on the willingness to respond

We found a significant main effect of the voice assistant's Elicitation Strategy: people perceived the voice assistant to be less disruptive when the voice assistant gives instructions on how to provide feedback (M = 1.93 , SD = 1.25), compared to when the voice assistant directly asks a question (M = 2.24 , SD = 1.06; β = -0.40, SE = 0.09, t = -4.34 , p < 0.01**). However, we observed no significant effect of the Elicitation Strategy on the willingness to respond, the acceptable frequency, or perceived cognitive load to answer.

We found a small effect with a Cohen's d = 0.37 between the question and instruction Elicitation Strategy on the perceived disruptiveness. Figure 4 shows the differences in participants' willingness to respond and the perceived disruptiveness for the different strategies.

In our study, instructions by the voice assistant tend to be longer ($M_{Instruction}$ = 20.56 words $SD_{Instruction}$ = 6.65 words; $M_{Question}$ = 14.78 words; $SD_{Question}$ = 7.10 words), although the difference is not significant (t = 1.83, p = 0.09). Intuitively, longer requests plus the additional effort of the user having to provide the feedback, seem to create more experience friction. However, our results suggest the opposite. Despite these results, further studies are needed to learn how the length of a voice assistant's request for feedback affects people's perceptions irrespective of the elicitation strategy.

A potential explanation for the lower disruptiveness of instructions is that it might feel less forced and more polite to be given the choice to provide feedback. However, with a choice, there

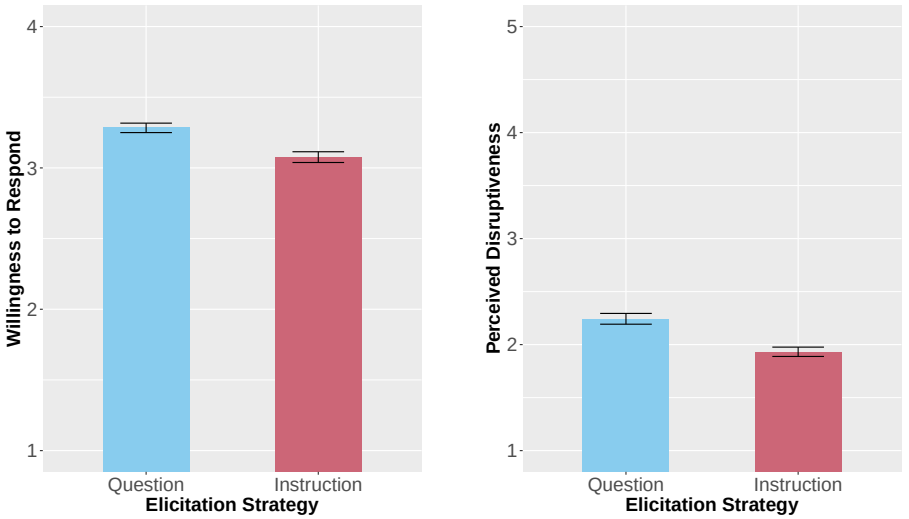

Fig. 4. Differences in participants' willingness to respond (4-point Likert Scale) and the perceived disruptiveness (5-point Likert Scale) for different Elicitation Strategies in our scenarios. People perceived the voice assistant to be less disruptive when the voice assistant gives instructions instead of direct question. However, we found no significant effect of the Elicitation Strategy on the willingness to respond.

is always the risk that the user will not respond. However, in our results, we did not observe a significant difference in users' willingness to respond. Therefore, providing instructions may elicit a similar amount of user feedback while avoiding too much perceived disruptiveness.

6.6 Elicitation Strategies become more important with higher levels of proactivity

As defined in the Study Design Section 5.2.3, each scenario we tested represented one of three levels of a voice assistant's proactivity. We found that although higher Levels of Proactivity reduced participants' willingness to respond (Low: M = 3.30, SD = 0.81; Mid: M = 3.10, SD = 088; High: M = 3.10, SD = 0.92) and increased the perceived disruptiveness (Low: M = 1.99, SD = 1.11; Mid: M = 2.24, SD = 1.20; High: M = 2.10, SD = 1.17), those differences are not significant (Willingness: β = -0.31, SE = 0.31, t = 0.59, p = 0.43; Disruptiveness: β = 0.07, SE = 0.08, t = 0.94 , p = 0.35).

We further looked at the interaction effects between Level of Proactivity and Framing, and between Level of Proactivity and Elicitation Strategy. We found a significant interaction effect of the latter on the perceived disruptiveness (t = -3.247, p < 0.01**): when the Level of Proactivity increases, so does the difference in perceived disruptiveness for the two Elicitation Strategies. Figure 5 shows this effect on the difference in the perceived disruptiveness for the two elicitation strategies.

Thus, the Elicitation Strategy plays a more important role when the feedback prompt is separate from an expected response to a user's request to the assistant. Scenarios where the voice assistant is asking for user feedback on a specific feature, or when trying to clarify behavioral signals, are common of the highest level of proactivity. For example, if a user leaves the music on for a long time and then the voice assistant initiates a new conversation to ask if the user is still listening, that would be considered a high level of proactivity. For those cases, our results suggest that instead of asking directly, the voice assistant should strongly consider providing instruction to avoid being too disruptive, e.g.,

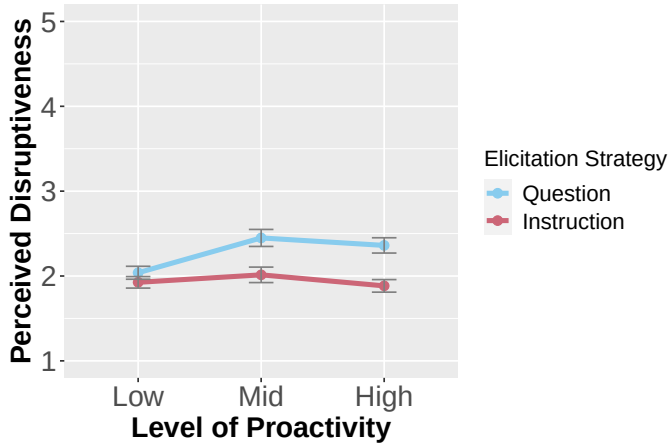

Fig. 5. Impact of Level of Proactivity on perceived disruptiveness for different Elicitation Strategies in our scenarios. Instruction shows a stronger effect on perceived disruptiveness with higher levels of proactivity.

"Hey just checking in. If you're still listening, say, 'Hey VA, I'm still listening.'"

6.7 A positive prior attitude towards voice assistants decreases the perceived disruptiveness

While not part of our main research question, we also found that participants who had a more positive attitude towards the voice assistant prior to the study perceived voice assistant prompts as less disruptive (β = - 0.38, SE = 0.09, t = -4.32 , p < 0.01**), and more acceptable to be asked frequently (β = - 0.26, SE = 0.11, t = 2.29 , p < 0.05*). Participants who perceived voice assistants as more useful (β = - 0.30, SE = 0.14, t = -2.25 , p < 0.05*) and trustworthy (β = - 0.12, SE = 0.06, t = -2.01 , p < 0.05*), felt that the voice assistant's prompts were easier to answer.

7 DISCUSSION

Our results show that we can influence users' willingness to respond to voice assistant requests for explicit user feedback and mitigate potential experience friction through intentional design choices. In this section, we discuss design implications and potential future directions for considering voice assistants as mediators between end users and the people building them.

7.1 Optimizing elicitation strategies for context and incentive

Prior to deciding on an elicitation strategy for the voice assistant's prompt, the user's general context should be considered. Voice assistants are ubiquitous now: people can interact with them on their phones, their headphones, their smart home devices, in their cars, and other places. This ubiquity provides rich opportunities for a voice assistant to collect user feedback in various situations, but also brings a new challenge.

As mentioned in [67], the interaction design of a voice assistant should account for various social settings in which the user's request and the voice assistant's response should be appropriate. For example, when the user is in a social setting with others nearby, probing for sensitive data, such as mood, may be inappropriate. People may consider such an interaction as a privacy threat, since other people on the scene may overhear the conversation [23]. To mitigate a user's privacy concern, one thread of research suggests giving the user control [16, 44]. The system should let the user

decide what information to give and when to give it. In our case, this implies that for feedback requests that may contain sensitive data, the safer design decision would be the elicitation strategy of providing instructions on how to give feedback whenever the user deems the moment more appropriate.

Our results suggest that people perceived the voice assistant's instructions to provide feedback as less disruptive compared to directly asking questions, without a significant difference in terms of participants' willingness to respond. However, the expected value of the elicited feedback should be considered for both the machine learning practitioner and for the user. If it is deemed crucial to improve or fix an underlying model, the strategy of asking a targeted question could help to more efficiently collect the most relevant user feedback even if it risks more disruption.

Of course, people are more encouraged to provide feedback if there is an incentive [12]. In some scenarios, the incentive can be implicit; for example, when the voice assistant is asking for input clarification for an ambiguous request, the user will immediately be rewarded with a better result. But when the voice assistant is probing for context, it might not be clear whether and how the collected information will benefit the user. Thus, including information on how it will benefit the user, or even offering other incentives, such as a monetary reward or free services, could likely increase a user's willingness to respond further.

7.2 Mediating collaborations in recommender systems

Building a recommender system can be described as a collaborative process between the stakeholders who build these systems (e.g., machine learning engineers, designers, and researchers) and millions of end users. Each stakeholder plays a different role in this collaboration. End users interact with recommender systems and leave traces of behavior that can be used by machine learning practitioners to build models. Researchers and designers run user studies to collect feedback about these systems to improve the user experience of these systems. In our study, we found people are willing to respond to a voice assistant's feedback requests, and perceive the interaction as less disruptive, when the voice assistant is framed as a collaborator. We argue we can further reinforce a voice assistant's collaborator role by letting a voice assistant connect the users of these systems with those who design and build them.

For example, user researchers and designers can work with machine learning engineers to design the types of voice assistant interactions that can help elicit feedback in a more enjoyable way. Once the feedback is collected, the voice assistant can then communicate the impact of the feedback to the user. This virtuous cycle can highlight users' contributions in the collaboration, which ultimately encourages them to provide more high-quality explicit feedback.

We can further extend the collaborative role of a voice assistant as a mediator, supporting the mutual interest between the system builders and users in improving the user experience. The voice assistant not only can collect feedback that a system builder needs, but also open the loop for users to advocate for their needs and ideas directly to the system builder. By opening the loop, the voice of a user could be directly heard by the system builder, which encourages broader participation in product design and development.

7.3 Taking advantage of the long-term and evolving relationship with an assistant

An important consideration for the design of voice assistants is that a user's experience is not shaped by a single interaction [10], and small language cues used by a voice assistant may play key roles in the user's perception [11]. Based on our results, one specific design implication would be to consistently reinforce the collaborative framing of the assistant. This would mean not only introducing it as a collaborator, but also having the assistant act accordingly in other interactions, using inclusive terminology like "we" and "let's" wherever appropriate to reinforce a collaborative

relationship. Such framing would also be reinforced by other design dimensions of a voice assistant such as voice and physical appearance. For example, the device that hosts the voice assistant could be designed like a student with a child's voice to reinforce the learner framing.

Although framing a voice assistant as a collaborator or a learner may be effective in eliciting user feedback, we should be careful to manage user expectations between the voice assistant's actual capability and the chosen framing. The discrepancy between the user's high expectations and the voice assistant's limited capabilities may backfire, resulting in a lower quality user experience [53, 54]. In this case, we could take advantage of the evolving relationship and consider a framing with low competence (e.g., learner) and gradually transition to high competence framing, such as a collaborator when the voice assistant becomes more capable.

Another opportunity of users' repeated engagement with a voice assistant is that it allows for a longitudinal approach to feedback elicitation. Feedback can be collected repeatedly over a longer period of time and lead to insights about changes in a user. This idea came up several times in our interviews, with the machine learning practitioners stressing that explicit in situ user feedback is particularly valuable when it captures change over time. Prior work on recommendation models also highlights the importance of incorporating the temporal change of a user's preference [41]. In such scenarios, the elicitation strategy and its design could adapt over time: as the user becomes more familiar with the ways the voice assistant asks for feedback, its prompts could be adjusted. For example, instructions could become shorter, or they could let the user know that feedback on this feature would help it improve its capabilities.

Overall, we found that a user's positive pre-existing attitude towards the voice assistant positively correlates with their willingness to answer a voice assistant's prompt and negatively correlates with the perceived disruptiveness. This suggests it is most effective to focus on creating the best possible experience in the first interactions with the assistant, and start asking for feedback only later in the relationship, when one can benefit from the positive impact of the earlier experiences. We could also consider a foot-in-the-door approach, in which a voice assistant starts with small, infrequent feedback requests for insensitive information (e.g., clarifying user input) and asks for more (e.g., understanding user context) later.

It also suggests that as voice assistants become more reliable and intuitive, user attitudes towards their capabilities may improve. Thus, we believe in the potential opportunities for explicit feedback elicitation as a core part of the voice user experience.

8 CONCLUSION

From reviewing related work and interviewing machine learning experts, we derived a set of four different usage categories where eliciting explicit feedback in situ on voice assistants would be valuable to improve the underlying user models and recommender systems powering today's voice assistants. We tested different design approaches, namely the Framing of the voice assistant, its Elicitation Strategy, and the Level of Proactivity in different scenarios, to better understand the influence that such decisions have on users' willingness to respond and the perceived disruptiveness in their experience. Our findings indicate that framing a voice assistant as a learning or collaborating entity instead of just an assistant can positively affect the users' perception, and that providing users with instructions on how to give feedback is perceived as less disruptive than asking direct questions. We also discuss design implications of these learnings and lay out future voice assistant design directions. The goal would be a future in which voice assistants could enable a seamless, collaborative process between end users and machine learning practitioners to further improve and personalize the assistants' services.

ACKNOWLEDGMENTS

We would like to thank Nediyana Daskalova, Jean Garcia-Gathright, and Brianna Richardson for their helpful comments. We appreciate the valuable input from our study participants, and the constructive feedback from our reviewers.

REFERENCES

[1] Noura Abdi, Kopo M Ramokapane, and Jose M Such. 2019. More than smart speakers: security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*.

[2] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*. Springer, 304–307.

[4] Gediminas Adomavicius and Alexander Tuzhilin. 2011. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.

[5] Miguel N Alexiades. 2003. Ethnobotany in the third millennium: expectations and unresolved issues. *Delpinoa* 45, 1 (2003), 15–28.

[6] Xavier Amatriain, Josep M Pujol, and Nuria Oliver. 2009. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 247–258.

[7] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[8] Anja Bachmann, Robert Zetzsche, Andrea Schankin, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. ESMAC: a web-based configurator for context-aware experience sampling apps in ambulatory assessment. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*. 15–18.

[9] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl-Heinz Lüke, and Roland Schwaiger. 2011. Incarmusic: Context-aware music recommendations in a car. In *International conference on electronic commerce and web technologies*. Springer, 89–100.

[10] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[11] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[12] Joost Broekens, Alina Pommeranz, Pascal Wiggers, and Catholijn M Jonker. 2010. Factors influencing user motivation for giving online preference feedback. In *5th Multidisciplinary Workshop on Advances in Preference Handling (MPREF'10)*.

[13] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.

[14] Julia Cambre, Ying Liu, Rebecca E Taylor, and Chinmay Kulkarni. 2019. Vitro: Designing a Voice Assistant for the Scientific Lab Workplace. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 1531–1542.

[15] Eugene Cho. 2019. Hey Google, Can I Ask You Something in Private?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[16] Eugene Cho, S Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. 2020. Will deleting history make alexa more trustworthy? effects of privacy and content customization on user experience of smart speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[17] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design*. Addison-Wesley Professional.

[18] Frederick G Conrad, Mick P Couper, Roger Tourangeau, Mirta Galesic, and Ting Yan. 2005. Interactive feedback can improve the quality of responses in web surveys. In *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.* 3835–3840.

[19] 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*. 1–12.

[20] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*. Springer, 429–446.

[21] JCF De Winter, Miltos Kyriakidis, Dimitra Dodou, and Riender Happee. 2015. Using CrowdFlower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing* 3 (2015), 2518–2525.

[22] Robert F DeVellis. 2016. *Scale development: Theory and applications*. Vol. 26. Sage publications.

[23] Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction* 31, 4 (2015), 307–335.

[24] Hauke Egermann, Mary Elizabeth Sutherland, Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. 2011. Does music listening in a social context alter experience? A physiological and psychological perspective on emotion. *Musicae Scientiae* 15, 3 (2011), 307–323.

[25] Mark G Ehrhart, Karen Holcombe Ehrhart, Scott C Roesch, Beth G Chung-Herrera, Kristy Nadler, and Kelsey Bradshaw. 2009. Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and individual Differences* 47, 8 (2009), 900–905.

[26] Julian J Faraway. 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

[27] Jon Froehlich, Mike Y Chen, Sunny Consolvo, Beverly Harrison, and James A Landay. 2007. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*. 57–70.

[28] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*. 2625–2633.

[30] F Maxwell Harper, Xin Li, Yan Chen, and Joseph A Konstan. 2005. An economic model of user rating in an online recommender system. In *International conference on user modeling*. Springer, 307–316.

[31] Daniel R Ilgen, Cynthia D Fisher, and M Susan Taylor. 1979. Consequences of individual feedback on behavior in organizations. *Journal of applied psychology* 64, 4 (1979), 349.

[32] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.

[33] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*. 47–51.

[34] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[35] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 229–237.

[36] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey Hancock, and Michael Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *arXiv preprint arXiv:2008.02311* (2020).

[37] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for In-vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–22.

[38] Jin Young Kim, Jaime Teevan, and Nick Craswell. 2016. Explicit in situ user feedback for web search results. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 829–832.

[39] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing*. Springer, 176–183.

[40] Sei Jin Ko, Charles M Judd, and Irene V Blair. 2006. What the voice reveals: Within-and between-category stereotyping on the basis of voice. *Personality and Social Psychology Bulletin* 32, 6 (2006), 806–819.

[41] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 447–456.

[42] I Han Kuo, Joel Marcus Rabindran, Elizabeth Broadbent, Yong In Lee, Ngaire Kerse, RMQ Stafford, and Bruce A MacDonald. 2009. Age and gender factors in user acceptance of healthcare robots. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 214–219.

[43] David R Large, Leigh Clark, Gary Burnett, Kyle Harrington, Jacob Luton, Peter Thomas, and Pete Bennett. 2019. " It's small talk, jim, but not as we know it." engendering trust through human-agent conversation in an autonomous, self-driving car. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–7.

[44] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.
[45] Song Liao, Christin Wilson, Long Cheng, Hongxin Hu, and Huixing Deng. 2020. Measuring the effectiveness of privacy policies for voice assistant applications. In *Annual Computer Security Applications Conference*. 856–869.
[46] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: a survey. *Decision Support Systems* 74 (2015), 12–32.
[47] 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 (2014), 94–100.
[48] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288
[49] Walid Maalej, Hans-Jörg Happel, and Asarnusch Rashid. 2009. When users become collaborators: towards continuous and context-aware user input. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*. 981–990.
[50] Alexander Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana, and Matthias Söllner. 2019. AI-based digital assistants. *Business & Information Systems Engineering* 61, 4 (2019), 535–544.
[51] Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N. Patel, and Jacob O. Wobbrock. 2015. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2953–2962. https://doi.org/10.1145/2702123.2702539
[52] Benjamin M Marlin. 2004. Modeling user rating profiles for collaborative filtering. In *Advances in neural information processing systems*. 627–634.
[53] Roger K Moore. 2017. Appropriate voices for artefacts: some key insights. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*.
[54] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer, 281–291.
[55] Michael G Morris and Viswanath Venkatesh. 2000. Age differences in technology adoption decisions: Implications for a changing work force. *Personnel psychology* 53, 2 (2000), 375–403.
[56] Chelsea M. Myers, Anushay Furqan, and Jichen Zhu. 2019. The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300277
[57] Clifford Nass, BJ Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678.
[58] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.
[59] David Nichols. 1998. Implicit rating and filtering. ERCIM.
[60] Derry O'Sullivan, Barry Smyth, and David Wilson. 2003. Explicit vs implicit profiling-a case-study in electronic programme guides. In *IJCAI*. 1351.
[61] Jeanne Parson, Daniela Braga, Michael Tjalve, and Jieun Oh. 2013. Evaluating voice quality and speech synthesis using crowdsourcing. In *International Conference on Text, Speech and Dialogue*. Springer, 233–240.
[62] Shwetak N Patel, Julie A Kientz, Gillian R Hayes, Sooraj Bhat, and Gregory D Abowd. 2006. Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In *International conference on ubiquitous computing*. Springer, 123–140.
[63] Cathy Pearl. 2016. *Designing voice user interfaces: principles of conversational experiences.* " O'Reilly Media, Inc.".
[64] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. In *Emotions and personality in personalized services*. Springer, 141–161.
[65] Štefan Pero and Tomáš Horváth. 2013. Opinion-driven matrix factorization for rating prediction. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 1–13.
[66] Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M Jonker. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 357–397.
[67] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[68] Stephen R Porter. 2004. Overcoming survey research problems. (2004).
[69] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.
[70] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 281–290.
[71] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 175–186.
[72] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180.
[73] Gale M Sinatra, CarolAnne M Kardash, Gita Taasoobshirazi, and Doug Lombardi. 2012. Promoting attitude change and expressed willingness to take action toward climate change in college students. *Instructional Science* 40, 1 (2012), 1–17.
[74] Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. 2018. The Ethnobot: Gathering Ethnographies in the Age of IoT. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[75] Donghua Tao. 2009. Intention to use and actual use of electronic information resources: further exploring Technology Acceptance Model (TAM). In *AMIA Annual Symposium Proceedings*, Vol. 2009. American Medical Informatics Association, 629.
[76] Jenn Thom, Angela Nazarian, Ruth Brillman, Henriette Cramer, and Sarah Mennicken. 2020. "Play Music": User Motivations and Expectations for Non-Specific Voice Queries. (2020).
[77] Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*. Springer, 479–510.
[78] David Traum. 2017. Computational approaches to dialogue. *The Routledge Handbook of Language and Dialogue. Taylor & Francis* (2017), 143–161.
[79] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 391–396.
[80] Jinke D Van Der Laan, Adriaan Heino, and Dick De Waard. 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies* 5, 1 (1997), 1–10.
[81] Chris Van Pelt and Alex Sorokin. 2012. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 765–766.
[82] Viswanath Venkatesh, Michael G Morris, and Phillip L Ackerman. 2000. A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational behavior and human decision processes* 83, 1 (2000), 33–60.
[83] Robert A Virzi. 1992. Refining the test phase of usability evaluation: How many subjects is enough? *Human factors* 34, 4 (1992), 457–468.
[84] Jordan Wirfs-Brock, Sarah Mennicken, and Jennifer Thom. 2020. Giving Voice to Silent Data: Designing with Personal Music Listening History. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.
[85] Jun Xiao, Richard Catrambone, and John T Stasko. 2003. *Be quiet? evaluating proactive and reactive user interface assistants*. Technical Report. Georgia Institute of Technology.
[86] Ziang Xiao, Michelle X Zhou, and Wai-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 437–447.
[87] 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 (2020), 1–37.
[88] Markus Zanker and Markus Jessenitschnig. 2009. Collaborative feature-combination recommender exploiting explicit and implicit user feedback. In *2009 IEEE Conference on Commerce and Enterprise Computing*. IEEE, 49–56.
[89] Yunfeng Zhang, Q Vera Liao, and Biplav Srivastava. 2018. Towards an optimal dialog strategy for information retrieval using both open-and close-ended questions. In *23rd International Conference on Intelligent User Interfaces*. 365–369.
[90] Kai Zheng, David A Hanauer, Nadir Weibel, and Zia Agha. 2015. Computational ethnography: automated and unobtrusive means for collecting data in situ for human–computer interaction evaluation studies. In *Cognitive informatics for biomedicine*. Springer, 111–140.
[91] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 425–434.

Received October 2020; revised April 2021; accepted July 2021