# MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction

Yucheng Jin
Lenovo Research
Beijing, China
jinyc2@lenovo.com

Wanling Cai
Department of Computer Science,
Hong Kong Baptist University
Hong Kong, China
cswlcai@comp.hkbu.edu.hk

Li Chen
Department of Computer Science,
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Nyi Nyi Htun
Department of Computer Science,
KU Leuven
Leuven, Belgium
nyinyi.htun@cs.kuleuven.be

Katrien Verbert
Department of Computer Science,
KU Leuven
Leuven, Belgium
katrien.verbert@cs.kuleuven.be

## ABSTRACT

Critiquing-based recommender systems aim to elicit more accurate user preferences from users' feedback toward recommendations. However, systems using a graphical user interface (GUI) limit the way that users can critique the recommendation. With the rise of chatbots in many application domains, they have been regarded as an ideal platform to build critiquing-based recommender systems. Therefore, we present *MusicBot*, a chatbot for music recommendations, featured with two typical critiquing techniques, user-initiated critiquing (UC) and system-suggested critiquing (SC). By conducting a within-subjects (N=45) study with two typical scenarios of music listening, we compared a system of only having UC with a hybrid critiquing system that combines SC with UC. Furthermore, we analyzed the effects of four personal characteristics, *musical sophistication (MS), desire for control (DFC), chatbot experience (CE), and tech savviness (TS)*, on the user's perception and interaction of the recommendation in *MusicBot*. In general, compared with UC, SC yields higher perceived diversity and efficiency in looking for songs; combining UC and SC tends to increase user engagement. Both MS and DFC positively influence several key user experience (UX) metrics of *MusicBot* such as interest matching, perceived controllability, and intent to provide feedback.

## CCS CONCEPTS

• **Human-centered computing → User interface design**; **Empirical studies in interaction design**.

## KEYWORDS

Critiquing-based recommender systems; conversational user interface; music recommendations; speech interaction; chatbot


*CIKM '19, November 3–7, 2019, Beijing, China*

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357923

**ACM Reference Format:**
Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357923

## 1 INTRODUCTION

A recent report by Grand View Research [1] predicts that the global chatbot market will reach $1.25 billion by 2025. Chatbots have utilized the power of artificial intelligence (AI) for various application domains ranging from customer services [12, 40] and healthcare [2, 23] to product recommendations [14, 43]. In the domain of recommender systems, there are several cases where product recommendations are delivered to customers through chatbots [14, 40] with an aim to improve customer engagement. At the same time, a number of research work [18–20, 22] have emphasized the importance of user control in recommender system.

Various studies with critiquing-based recommender systems (CBRS) [7, 10, 26] have shown the positive effects of increased interactivity on the effectiveness of recommendations. Critiquing is an iterative approach of evaluating the outputs of a recommender system, which allows the system to continuously update the settings and provide users with recommendations that better represent desired outcomes [10]. Figure 1 shows a typical interaction flow of CBRS. CBRS simulate an artificial salesperson who first recommends products based on a user's initial preferences and then shows a new set of products based on the user's feedback (aka critiques), e.g., "something cheaper", "larger screen", etc. Thus, CBRS are well suited to accommodate user control during the recommendation process.

Most existing research studies [9, 10] have compared different critiquing techniques with graphical user interfaces (GUIs). However, little work has studied different critiquing techniques with conversational user interfaces (CUIs) that mimic a conversation with a real human either by text or voice. Moreover, it has been shown that personal characteristics such as musical sophistication affect user perception of controls for music recommenders [20];

[1] https://www.grandviewresearch.com/press-release/global-chatbot-market

however, the effects of personal characteristics have not been validated on critiquing techniques yet. To fill these research gaps, this paper **compares two typical critiquing techniques with CUIs** and investigates **how personal characteristics influences user perception and interaction of recommended items** (see the dashed lines in figure 1). To achieve these objectives, we implemented a hybrid critiquing-based music recommender *MusicBot*, which uses a chatbot to enable users to interact with recommendations through both text and voice. The system offers two major critiquing techniques, user-initiated critiquing (UC) and system-suggested critiquing (SC) to refine the recommendation. UC enables users to construct critiques according to their own needs, while SC generates a set of critiquing candidates for users to choose a desired critique. We then conducted an evaluation with 45 participants using *MusicBot* in a within-subject design.

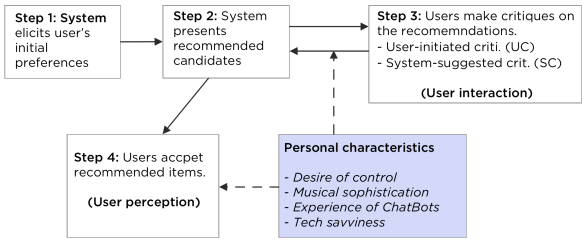


**Figure 1: A typical interaction flow of critiquing-based recommender systems. The relation between personal characteristics (PC) and Steps 3 & 4 shows the potential effect of PC on user perception of, and interaction with, recommended items.**

We raise three research questions for evaluating critiquing-based Music recommenders **particularly with a conversational user interface (CUI)**.

**RQ1**: Which critiquing setting, UC versus HC, is better suited for controlling music recommendations?

**RQ2**: Which personal characteristics (e.g. musical sophistication, desire for control, chatbot experience, and tech savviness) might influence user's perception and interaction of recommendations?

**RQ3**: Are critiquing techniques perceived as useful in low-involvement product domains as in high-involvement product domains?

Our main contributions are four-fold:

(1) We demonstrate a multi-modal (text and voice) conversational music recommender that incorporates both a user-initiated critiquing technique (UC) and a system-suggested critiquing technique (SC). We then employ a mixed qualitative and quantitative research method to compare UC with a hybrid critiquing technique (HC) in terms of subjective user experience (UX) with recommendations. Overall, recommendations generated by UC and HC were perceived at the same level, while users tend to need more effort to find a song using HC.

(2) We find that two personal characteristics, *desire for control* and *musical sophistication*, positively influence several key UX metrics of recommendations such as interest matching, intent to give feedback, and perceived controllability.

(3) Our study also verified the usefulness of critiquing techniques in a low-involvement domain of music recommendations.

(4) Based on the findings in this study, we proposed specific design suggestions for critiquing-based recommender system with conversational interaction.

This paper is organized as follows: We first introduce related work, followed by the design and implementation of *MusicBot*. We then present the quantitative and qualitative results of a user study. Finally, we conclude with a discussion of study findings and limitations.

## 2 RELATED WORK

In the following sub-sections, we review previous work that are closely related to our research.

### 2.1 Critiquing-based Recommendations

Based on the way of constructing critiques, critiquing-based recommender systems (CBRS) can be categorized into two types of critiquing: system-suggested, and user-initiated. In addition, the distinction is made between *unit critiques* and *compound critiques*. *Unit critiques* refer to critiques that only constrain a single feature at a time, while *compound critiques* are capable of making a critique over multiple features simultaneously to improve performance of conversational recommender systems[26]. Due to the pros and cons of each type of critiquing technique [10], by taking the advantages of both UC and SC, a hybrid system increases decision accuracy and needs less cognitive effort[7]. However, most studies [10] of comparing different critiquing techniques are conducted only with graphical user interfaces (GUIs). To enable critiques with conversational interaction, we evaluate a hybrid critiquing system in a multi-modal chatbot for music recommendations.

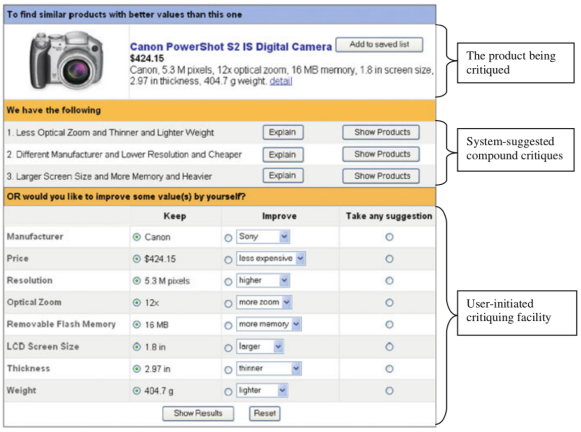


**Figure 2: A user interface of a hybrid critiquing system that combines UC and SC [7].**

**User-Initiated Critiquing (UC)**. UC is a flexible critiquing approach that allows users to determine which features and how the features are critiqued (see Figure 2). Thus, users are able to make either unit critiques or compound critiques. This technique is particularly useful for tradeoff navigation between compromising values

on less important attributes and obtaining more optimal values for important attributes. The most representative systems of UC are *Example Critiquing* [31] and *Flat Finder* [38]. UC empowers users to have a higher level of user control, which does not lead to higher perceived cognitive load. However, some previous user studies [10] of UC also found users may suffer from the difficulty of getting started with UC without prior knowledge.

**System-Suggested Critiquing (SC)**. Instead of asking users to construct critiques, SC generates a set of critique candidates for users to pick (see figure 2). Generating critiques is based on the system's knowledge about the current user's preference and the availability of remaining products. The earlier systems of SC, *FindMe* [5] and *ATA* [24], presented pre-designed *unit critiques* to users with less adaptation to the changes of user preference and interaction. Later on, Reilly et al. [33] proposed *Dynamic Critiquing* based on association rules such as Apriori algorithm [1] to find frequent sets of value differences between the recommended product and the remaining alternatives. Furthermore, *Incremental Critiquing* [34] greatly reduces interaction cycles by avoiding to show the user rejected critiques in history. To take into account users' interest in the suggested critiques, Zhang and Pu [42] proposed to generate MAUT (Multi-Attribute Utility Theory) based compound critiques. The approach significantly increases recommendation quality by ranking the critique candidates based on the overall satisfaction degree with user preferences. A more advanced SC is the *Preference-based Organization* technique [8] that is able to generate more diversified compound critiques and achieve higher critique prediction accuracy and recommendation accuracy. Overall, SC is able to produce more dynamic critiques based on the current user's preferences. It is specially useful for users who have difficulties to initiate critiques or build critiques by themselves. However, domain experts may call for more control over recommender systems, so SC may restrict the way they intend to make critiques.

**Hybrid Critiquing (HC)**. Similar to the idea of hybrid recommender systems [6], HC intends to take advantage of each type of critiquing techniques. Chen and Pu [7] compared two hybrid critiquing systems that combine a UC system (*Example Critiquing*) with a SC system (*Dynamic Critiquing* or *Preference-based Organization*) in a graphical user interface. Users showed positive attitudes toward HC that comprises both UC and SC. In addition, HC can also overcome the issues of adopting a single type of critiquing technique and lead to high decision accuracy and low objective effort in making a choice.

All research findings discussed above were tested only with graphical user interfaces. In contrast, this study tries to compare different critiquing techniques with a conversational user interface.

## 2.2 Conversational Recommender Systems

Conversational interaction is well suited for critiquing the recommendation through natural language. Several works have demonstrated systems that elicit user preference and present recommendations in a conversational dialog. ExpertClerk [36] is a conversational agent that acts as a human salesclerk to *ask* user questions for getting user shopping preference and *proposes* the matched products with explanations. Adaptive place advisor [37] provides personalized recommendations for traveling places. The system refines user queries by considering both long-term interests over many conversations and short-term interests in the current conversation. The two systems are typing-based conversational recommender systems.

As voice recognition techniques continue to improve, speech interaction is becoming more capable of allowing users to express more complex content. ReComment [15] presents a speech-based user interface for making unit critiques (critiquing over a single feature at a time), and it improves the perceived ease of use as well as the overall quality of recommendations. A recent study [21] found that users tend to express longer and more conversational content with speech-based user interfaces than with typing-based user interfaces. However, speech user interfaces might negatively influence the efficiency of consuming recommendations and user exploration [41]. So far, most speech-based UIs for recommender systems are still featured with search-oriented commands rather than more sophisticated commands to control recommendations.

To the best of our knowledge, the existing critiquing systems with speech interaction only incorporate user-initiated critiquing. Little work has integrated system-suggested critiques into the dialog-based conversational recommender system. In addition, the effect of users' personal characteristics on their interaction behavior and subjective perception of critiquing-based systems has not been investigated yet.

## 2.3 Personal Characteristics

Although previous research [20, 28] has shown how personal characteristics influence the way users control music recommendations through an interactive visualization, we do not know whether personal characteristics also affect user perception and interaction of critiquing based recommendations. In the following paragraphs we explain the four personal characteristics we have considered in this paper as well as the rationale for selecting them.

**Desire for Control (DFC)**. Greenberger et al. [16] first used a questionnaire to measure DFC in various work-related tasks of their new jobs. Users with higher DFC tend to perform better on the task and do better on upcoming tasks than subjects with low DFC [3]. We use a widely used DFC scale proposed by Burger et al. [4] to measure the degree of control individuals perceive towards outcomes. DFC is an important personal characteristic (PC) to measure for this study, since the two different critiquing techniques in our system empower users to have different levels of user control.

**Musical Sophistication (MS)**. MS has been found as a key PC that influences the way users interact with music recommender systems [20]. The Goldsmiths Musical Sophistication Index (Gold-MSI) [29] is an effective test for measuring domain knowledge of participants. Several studies investigating the effect of personal characteristics on music recommender systems have employed the Gold-MSI to measure the participant's musical sophistication.

**Tech Savviness (TS)**. TS often reflects a participant's confidence in trying out new technology. Several studies [13, 27, 28] have investigated how TS may influence the way participants interact with recommender systems. Therefore, we think TS may also influence the way participants critique recommendations in a conversational agent.

**Chatbot Experience (CE)**. Due to the impact of assimilation bias, participants with previous chatbot experience are prone to overestimate or underestimate the sophistication of using a chatbot [11]. Previously, when conversational agents were not popular, researchers often measured participants' previous experience with computers as an influencing factor for conversational agents [2, 35]. A recent study [25] measures the effect of previous experience with voice user interfaces on a voice-based conversational agent. As chatbots are becoming pervasive in everyday life, we measure CE of participants in our study.

# 3 SYSTEM DESIGN AND IMPLEMENTATION

## 3.1 Work Flow

Figure 3 illustrates the working flow of *MusicBot*. *MusicBot* is a multi-modal chatbot that enables both text and voice for input and output. The working flow consists of seven steps: (1) A user sends a text/voice message to bot. (2) The web client transfers the message to a service for natural language understanding implemented by Dialogflow [2]. (3) The message is processed and matched to a corresponding intent. (Intents are defined beforehand by developers in Dialogflow and correspond to the supported critiques and control commands such as "next".) (4) When a certain intent is identified by Dialogflow, Dialogflow sends a formatted response as actionable data. (5) The actionable data corresponding to messages that go to the client directly. (6) The actionable data related to music selection calls the Spotify web API [3] to generate recommendations. (7) The user receives text/voice messages and music.

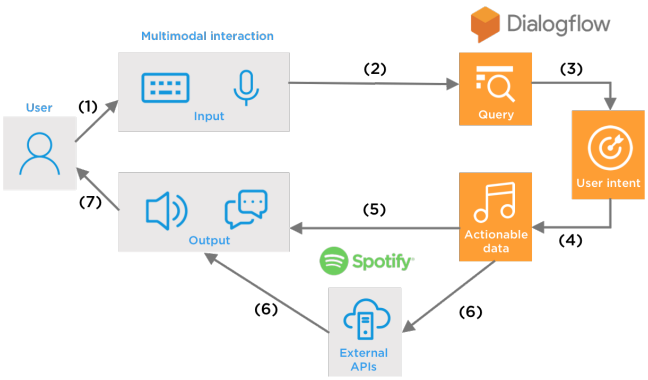


**Figure 3: The interaction work flow of *MusicBot*.**

## 3.2 Algorithm

*3.2.1 Recommendation Algorithm.* The Spotify API generates recommendations based on three types of seeds, i.e., songs, artists, and music genres. We first ask users to specify at least one item for each type of seeds. As the API provides an option to set the preferred range (low, medium, high) for five key audio features (danceability, speechiness, energy, valence, and tempo), we also ask users to indicate what kind of music they like in terms of a specific audio feature. Users can, for instance, set tempo to high for getting more "fast" music. After building the **user profile**, the system generates 50 songs based on each type of seeds. In total, 150 songs are generated and added to a playlist as initial recommendations for users.

*3.2.2 Critiquing-based Algorithm.* After receiving a recommendation, users are able to critique it by sending a text/voice message (UC) or asking for suggestions from the system by clicking the button "Let bot suggest" (SC) in our *MusicBot*. For instance, users can make critiques on music-related attributes (i.e., genre, language, artist, danceability, speechiness, energy, valence, and tempo) by themselves (e.g., *"I need lower energy"*) or request the system to give some suggestions (e.g., *"Based on your music preference, we think you might like English songs with higher danceability and higher energy."*). The elicited user preferences from UC and SC are used to dynamically update both the user profile and the current playlist. As for system-suggested critiques in our *MusicBot*, we adopt a heuristic method that was validated by [8] to generate personalized and diverse critiques on the current recommended item in terms of the above-mentioned music-related attributes. The generation of critique considers user preferences captured from the previous interaction, which involves four steps:

(1) We convert each candidate song in the current playlist to a critique pattern vector (e.g., {*(energy, higher), (danceability, similar), ..., (valence, higher)*}), by comparing it with the current recommended item with regard to each music-related attribute.
(2) We then select the frequent occurring subsets of critique pattern, representing the characteristics of music that users may prefer, from all critique pattern vectors via a popular associate rule mining algorithm (i.e., Apriori algorithm). These selected subsets may each contain two or three attributes, so they are also called compound critiques [26]. With these compound critiques, the current playlist can be grouped into different categories for subsequent recommendations.
(3) We calculate the utility of each selected compound critique [8] (according to multi-attribute utility theory (MAUT) [42]), based on the component of critiques as well as the predicted user preferences over their contained music.
(4) We finally select a set of personalized and diversified critiques to assist the user in seeking music, by multiplying the critique utility with a diversity degree as suggested in [8].

## 3.3 User Interface Design

Figure 4 shows the user interface designed for our study. The interface was designed to fulfill two requirements: 1) It should stimulate music search on a conversational user interface for the presented scenario, and 2) it should enable users to critique recommendations by user-initiated critiquing (UC) or system-suggested critiquing (SC). The interface consists of the following components: The *MusicBot* prototype, an instruction panel, and a rating widget. Below we describe each component in detail.

The designed *MusicBot* interface follows the current popular chatting platforms such as Messenger and WeChat. A dialog window (Figure 4, b) shows all happened conversations between the

[2] https://dialogflow.com
[3] https://api.spotify.com/v1/recommendations

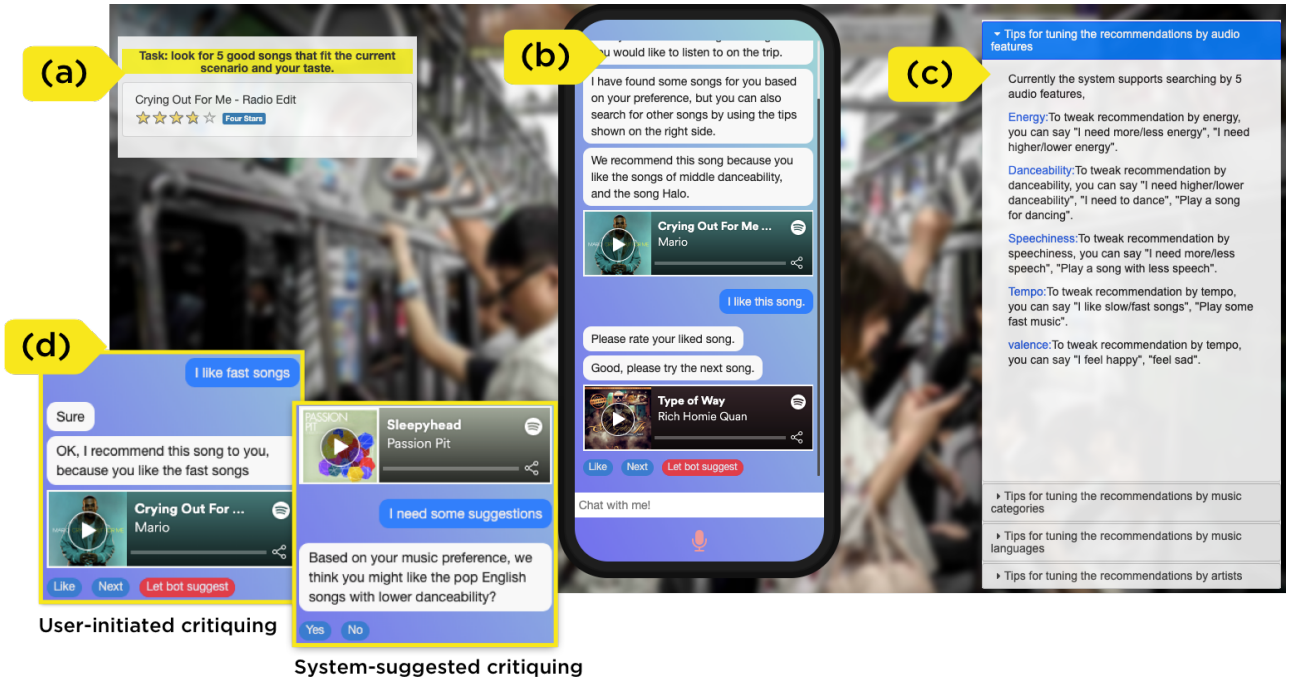


**Figure 4: The User Interface highlighting the scenario of "taking a subway after work"; a) a list for showing a user's liked songs with star rating, b) the prototype of *MusicBot*, c) an instruction panel showing the supported critiquing features and some examples, d) two examples showing the dialog flow for user-initiated critiquing and system-suggested critiquing respectively.**

user and the bot. Of note, the bot also speaks the content of a message when the message appears in the dialog. The bot shows a song in a card, which allows the user to control music. For each song, we show a set of feedback buttons under the card. Specifically, by clicking the "Like" button, the current played song will be added to the list of liked songs for user rating (Figure 4, a). The "Next" button allows the user to skip the current song and play the next song in the playlist. The "Let bot suggest" button (red) triggers system-suggested critiquing for the current song, and returns a set of critiques. The bottom part is an input panel for sending messages, either by typing or voice. To support the user in making user-initiated critiques, an instruction panel explains the supported critiquing features with some examples (Figure 4, c).

In addition, we show two examples of making **user-initiated critiques** (UC) and **system-suggested critiques** (SC) in dialogues (Figure 4, d). For UC, after understanding the user's intention to critique the current song on "tempo" with a higher value, the bot recommends a new song after explaining the recommendation. In the dialog of SC, the user first clicks "Let bot suggest", and then the bot shows one critique from a set of suggested critiques. The user will be asked to accept the current critique (Yes) or view the next critique (No).

## 4 EXPERIMENTAL DESIGN

We conducted a within-subjects user study (N=45) to compare a system only supporting UC with a hybrid system combining both UC and SC. Therefore, we disabled SC in *MusicBot* as the *baseline* condition. To minimize the learning effects, half of the participants evaluated two interfaces in a reverse order.

### 4.1 Participants

We initially recruited 51 participants through personal contacts, research groups, and university contacts for the study. On completion, users were allowed to enter a prize draw to win one out of 20 vouchers each worth 10 USD. Two participants' responses were removed because they did not finish the study within one hour, and four participants were rejected due to the low quality of their data. We finally kept the data of 45 participants (Age: 20-30(36), 30-40(6),41-50(1), > 50(2); Gender: Female = 19, Male = 26).

### 4.2 Procedure

The **experimental task** was to find five songs that best match the presented scenario and the music preference of the user. Each participant needed to perform this task for two different scenarios: Taking the subway after class/work, and organizing a friend's birthday party. The procedure contains the following steps:

(1) *Tutorial of study* - Participants were first asked to read the description of the user study and watch a video tutorial that introduces the main features of the system. Considering some participants are not Spotify users, we provided a public Spotify account to authorize to the system.
(2) *Building user profile* - Participants were asked to specify a sample of up to three of their favorite artists, songs and

music genres, as well as their preferences for five audio features (danceability, speechiness, energy, valence, and tempo). Explanation of each audio feature was also provided.

(3) *Pre-study questionnaire* - This questionnaire asked the participant's age and gender, and measured four personal characteristics (musical sophistication, desire for control, chatbot experience, and tech savviness).

(4) *Warming up* - Before starting the given task, participants were allowed to get familiar with the system by trying the supported features listed in the instruction panel.

(5) *Making critiques* - Based on the scenario and the task introduced in the conversation, participants needed to react to the recommended songs one by one either by accepting/skipping the current song or making a critique.

(6) *Post-study questionnaire* - Participants filled a post-study questionnaire after finishing the task in each scenario. Based on a user-centric evaluation framework for recommender systems [30], our questionnaire included 14 questions for measuring user perceptions of recommender systems. Users were able to provide free text comments in the end.

## 4.3 Materials

Participants were asked to fill a pre-study questionnaire and two post-study questionnaires for measuring user perceptions of recommendations in two scenarios.

The pre-study questionnaire contains questions for measuring four personal characteristics. We employed ten statements from the sub-scale "general MSI" of Goldsmiths Musical Sophistication Index (Gold-MSI)[4] to measure musical sophistication (MS) of the participant. For measuring desire for control (DFC), we used 20 statements proposed by Burger and Cooper [4]. The chatbot experience (CE) and tech savviness (TS) were measured by two statements: *"I often use a chatbot (such as Siri, Cortana) on my personal devices."* and *"I am confident when it comes to try new technology."* respectively. The first post-study questionnaire consists of 14 statements (Table 1), which are mainly based on the user-centric evaluation framework for recommender systems [31] to gauge how critiquing settings influence user experience; and the statements of Q10-Q12 are from an evaluation framework for conversational agents [39], which focus on user experience with conversation. In addition to the 14 statements, the second questionnaire includes one additional question to ask the user's preference for critiquing technique, and an open ended question for obtaining her/his free comments on the system.

All statements were measured by 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree).

## 4.4 Interaction Logs

We captured user conversations for analyzing their actual behavior during the interaction with *MusicBot*. Based on this log data, we then calculated the following metrics for both experimental scenarios:

- Rating: The average star rating for the five liked songs.
- Completion time: The time a participant spent on finishing a task.

[4]http://www.gold.ac.uk/music-mind-brain/gold-msi/

**Table 1: Post-study Questionnaire for Testing User Experience with *MusicBot*.**

| Question items |
|---|
| Q1: The items recommended to me matched my interests. |
| Q2: I easily found the songs I was looking for. |
| Q3: Looking for a song using this interface required too much effort (reverse scale). |
| Q4: The songs recommended to me are diverse. |
| Q5: I found it easy to inform the system if I dislike/like the recommended song. |
| Q6: I felt in control of modifying my taste using *MusicBot*. |
| Q7: I am confident I will like the songs recommended to me. |
| Q8: I like to give feedback on the music I am listening. |
| Q9: This music chatbot can be trusted. |
| Q10: I found the system easy to understand in this conversation. |
| Q11: In this conversation, I knew what I could say or do at each point of the dialog. |
| Q12: The system worked in the way I expected in this conversation. |
| Q13: I will use this music chatbot again. |
| Q14: Overall, I am satisfied with the chatbot. |

**Table 2: Descriptive Statistics for User Interaction Data. Significance: *** p < .001, ** p < .01, * p < .05.**

| Interaction metrics | UC (mean,sd) | HC (mean,sd) |
|---|---|---|
| Rating (stars) | (4.05, 0.47) | (4.08, 0.44) |
| Completion time* (minutes) | (5.40, 4.19) | (6.98, 4.16) |
| #Listened songs** | (10.67, 4.99) | (13.13, 6.09) |
| #Turns(times)** | (12.29, 8.21) | (16.11, 9.35) |
| #Btn(times)*** | (9.18, 3.38) | (12.64, 7.07) |
| #Typing(times) | (3.09, 4.78) | (3.07, 4.21) |
| #Voice(times) | (1.24, 7.90) | (0.71, 2.97) |
| #Words | (2.13, 1.92) | (2.28, 1.84) |
| #Unknown utterances | (1.78, 6.46) | (0.78, 1.80) |

- #Listened songs: The total number of songs listened by a participant before finishing a task.
- #Turns: The number of dialog turns before finishing a task.
- #Btn: The number of clicks on buttons.
- #Typing: The number of typed utterances.
- #Voice: The number of utterances sent by voice.
- #Words: The average number of words per utterances.
- #Unknown utterances: The number of utterances that were not understood correctly by the bot.

# 5 RESULTS

## 5.1 Subjective Experience

Figure 5 presents the results of users' responses to the statements shown in Table 1. We performed a non-parametric Mann-Whitney test to compare the two critiquing settings (UC and HC) regarding user perceptions of recommendations. The results only show a

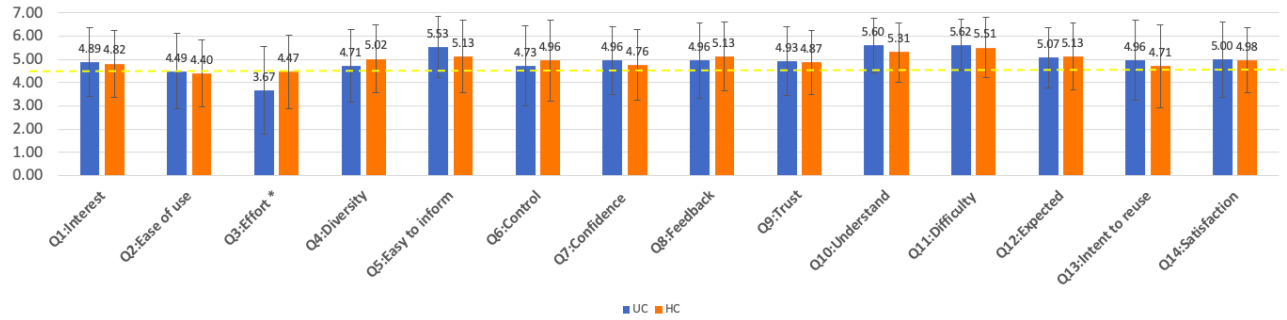


**Figure 5: Usability and user satisfaction assessment results. A cut off value at 4.5 represents agreement on the 7-point Likert scale. * is marked for significant difference at the 5% level (p-value < 0.05).**

significance between UC (Mean=3.72, SD=1.81) and HC (Mean=4.54, SD=1.55) on the *effort of looking for songs* (U = 919.500, p = .02). We also calculated the effect size by dividing $Z$ by the square root of the total number of the samples, which shows a medium effect ($r$ = 0.31) [5]. We do not find a significant difference between two systems on the remaining aspects. By setting a cut-off value at 4.5 for agreement on 7-point Likert scale, we find that users positively rated UC and HC in most of the UX metrics (Q1, and Q4-14) in the questionnaire.

Furthermore, we analyzed how the actual use of system-suggested critiques (SC) affects user perception of recommended songs in the condition of HC. Though there was no significance, we observe a clear trend that the users who tried SC (24 users) tend to perceive higher ease of use (Q2) and diversity (Q4) than those who did not try SC (21 users). However, using SC tends to increase user effort to look for songs (Q3). Interestingly, the users who did not try SC are more confident (Q7) in the liked songs and feel easy to give feedback (Q5) and control the system (Q6).

## 5.2 User Behavior

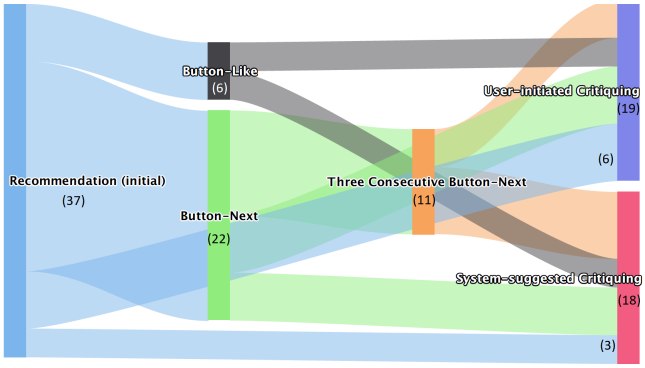


**Figure 6: Users' interaction flows in HC (starting from initial recommendation till users critique recommendations).**

We further analyzed users' log data to see they actually interacted with *MusicBot* using the two critiquing techniques UC and HC. The log recorded nine interaction metrics as shown in Section 4.4. We first ran the Shapiro-Wilk test for testing normality of data. We then employed t-tests for normally distributed data and Wilcoxon signed-rank tests for non-normally distributed data.

*5.2.1 Task Performance.* Table 2 shows that on average HC led to significantly more dialog turns (#Turns) than UC did in a task (t = -2.58, p=.007), which in turn led to significantly more time (Completion time) to finish a task in HC than in UC (t = -2.06, p=0.02). During the task, users tended to try significantly more pieces of music (#Listened songs) in HC than in UC (t = -2.56, p=.007).

In addition, our participants gave a relatively high rating for their liked songs: The average ratings of their liked songs in both systems were above 4 out of 5 stars. This may indicate that *MusicBot* can provide users with satisfying recommendations. To investigate which critiquing technique is more effective for looking for a song, we analyzed the provenance of the songs liked by users. On average, we find that 42.6% of liked songs were found after making UC and 24.4% after making SC, and the rest of the liked songs were from initial recommendations.

*5.2.2 Interaction Behavior.* We observed users' interaction behavior on different interaction modalities (i.e., button, typing, voice) in the two systems. On average, HC led to significantly more interaction with buttons than UC (t = -3.68, p<.001). However, we did not find any significant differences on the remaining interaction metrics.

In addition, to explore when users would use UC and SC for critiquing recommendations, we further analyzed the typical interaction flow which starts from initial recommendation till users made the first critique (SC/UC) on recommendations in HC. (see Figure 6). The results show that 82.22% (37/45) of our participants employed UC or SC when seeking for music recommendations, among which most participants (59.46%, 22/37) critiqued recommendations after they clicked the button "Next". Some of them (29.73%, 11/37) used UC or SC after consecutively clicking the button "Next" three times (users will receive a reminder of critiquing settings from the bot if they consecutively click the "Next" button three times). Moreover, 24.32% (9/37) of participants critiqued the initial recommendation and 16.21% (6/37) of them used UC or SC even if they received ideal recommendations.

[5] small effect: $0.1- < 0.3$, medium effect: $0.3- < 0.5$, large effect: $\geqslant 0.5$

**Table 3: The Effect of PC on Users' Perceptions of Recommendations measured by Pearson correlation coefficient. Significance: *** p < .001, ** p < .01, * p < .05.**

| PC | Q1:Interest | Q2:Ease of use | Q3:Effort | Q4:Diversity | Q5:Easy to inform | Q6:Control | Q7:Confidence |
|---|---|---|---|---|---|---|---|
| CE | 0.15 (0.33) | 0.14 (0.37) | 0.07 (0.66) | 0.03 (0.84) | -0.03 (0.86) | 0.11 (0.46) | 0.05 (0.73) |
| TS | -0.01 (0.98) | -0.13 (0.40) | **0.36 (0.02)*** | 0.10 (0.51) | -0.08 (0.59) | -0.19 (0.21) | -0.12 (0.43) |
| MS | **0.40 (0.01)*** | 0.25 (0.10) | -0.22 (0.14) | 0.17 (0.26) | 0.10 (0.53) | **0.31 (0.04)*** | 0.29 (0.05) |
| DFC | 0.23 (0.14) | 0.03 (0.84) | 0.13 (0.41) | 0.24 (0.11) | 0.22 (0.15) | **0.35 (0.02)*** | 0.25 (0.10) |

| PC | Q8:Feedback | Q9:Trust | Q10:Understand | Q11:Difficulty | Q12:Expected | Q13:Intent to reuse | Q14:Satisfaction |
|---|---|---|---|---|---|---|---|
| CE | 0.06 (0.70) | -0.01 (1.00) | -0.07 (0.65) | 0.02 (0.88) | 0.06 (0.69) | 0.21 (0.17) | 0.10 (0.52) |
| TS | 0.16 (0.29) | 0.07 (0.66) | -0.12 (0.42) | -0.04 (0.77) | 0.04 (0.78) | -0.12 (0.42) | -0.19 (0.10) |
| MS | **0.55 (<0.001)***** | **0.37 (0.01)*** | 0.09 (0.57) | 0.13 (0.38) | 0.23 (0.14) | **0.31 (0.04)*** | 0.22 (0.15) |
| DFC | 0.06 (0.68) | 0.16 (0.29) | **0.30 (0.04)*** | **0.38 (0.01)*** | 0.22 (0.14) | 0.28 (0.06) | 0.20 (0.19) |

**Table 4: The Effect of PC on Users' Interaction with Recommendations measured by Pearson correlation coefficient. Significance: *** p < .001, ** p < .01, * p < .05.**

| PC | #Listened songs | Rating | Completion time | #Turns | #Btn |
|---|---|---|---|---|---|
| CE | -0.04 (0.79) | 0.26 (0.08) | -0.05 (0.74) | -0.05 (0.76) | -0.11 (0.45) |
| TS | -0.11 (0.46) | 0.04 (0.79) | -0.25 (0.10) | -0.08 (0.59) | -0.10 (0.51) |
| MS | -0.14 (0.35) | **0.45 (0.002)**** | -0.17 (0.28) | -0.15 (0.34) | -0.15 (0.31) |
| DFC | 0.04 (0.79) | 0.12 (0.44) | 0.14 (0.35) | -0.004 (0.98) | -0.08 (0.58) |

| PC | # Typing | #Voice | # Words | #Unknown utterance |
|---|---|---|---|---|
| CE | -0.09 (0.56) | 0.25 (0.10) | 0.22 (0.15) | 0.02 (0.90) |
| TS | -0.07 (0.67) | 0.06 (0.72) | 0.13 (0.38) | -0.02 (0.89) |
| MS | -0.02 (0.90) | -0.05 (0.73) | 0.14 (0.36) | 0.04 (0.81) |
| DFC | 0.15 (0.34) | -0.001 (1.00) | 0.11 (0.46) | 0.12 (0.42) |

Besides, we calculated the conversation rate of each critiquing technique by counting the ratio of the number of accepted songs to the number of performed critiques. The conversation rate of SC (45.10%) is much higher than that of UC (28.48%), implying that it may be more helpful for users to find a satisfactory song right after the critique.

## 5.3 Personal Characteristics

Since we were particularly interested in investigating the effect of PC in a condition that has both UC and SC, we only considered the data collected from the hybrid critiquing system. We performed a Pearson Correlation analysis to understand how PC influences user perceptions of, and interaction with, the recommended songs.

*5.3.1 Correlation between PC and User Perception.* Table 3 shows all correlations between the four personal characteristics (CE, TS, MS, and DFC) and different UX aspects measured by statements as shown in Table 1. The significant correlations are presented in boldface. Specifically, we find that tech savviness (TS) is positively correlated with the effort of looking for a song (r = 0.36, p<.05). Musical sophistication (MS) is *positively* related to interest matching (r = 0.40, p<.01), controllability (r = 0.31, p<.05), intention to give feedback (r = 0.55, p<.001), trust (r = 0.37, p<.05), and intent to reuse (r = 0.31, p<.05). Moreover, the desire for control (DFC) *positively* influenced multiple UX metrics including controllability (r = 0.35, p<.05), easy to understand (r = 0.30, p<.05), and knowing how to critique (r = 0.38, p<.01). To view more detailed correlation coefficients, please refer to the Table 3. We did not find significant correlations between Chatbot experience (CE) and any UX aspects.

*5.3.2 Correlation between PC and User Interaction.* Table 4 shows the correlations between four PCs and nine interaction metrics, but we only find that MS has a strongly positive correlation with user ratings of the liked songs (r = 0.45, p<.01).

Meanwhile, we performed a moderation analysis to investigate whether the four personal characteristics moderate the significant effect of critiquing settings on user effort of looking for songs. However, we did not find a significant moderating effect for any PC.

## 5.4 Subjective Feedback

The second post-study questionnaire also includes one question asking participants to indicate which critiquing technique (UC versus SC) they prefer and to explain why they prefer this critiquing

technique, along with an open question used to collect other comments related to *MusicBot*. We performed a correlation analysis between four personal characteristics and user-indicated preference for critiquing technique, but we did not find any significant correlation. The main reasons for choosing UC over SC are: (1) Users were confident in finding good songs through the critiques made by themselves; and (2) they have higher demand for controllability. Users who prefer SC indicate that: (1) Sometimes they do not know what kind of music they want to listen to; (2) they get tired of constructing the critiques; and (3) they like to be surprised and discovering news songs. Overall, participants commented positively about *MusicBot*, e.g., *"I really like the concept of a chatbot suggesting (new) music to me... (P18)"*. and *"fantastic work! it's not a commercial application yet and imo the chatbot does not handle simple nlp problems, but I enjoyed a lot testing it and interacting with it... (P35)"*.

## 6 DISCUSSION

To answer the raised research questions, we discuss how the critiquing settings and personal characteristics influence the perceived quality of recommendations and the way users interact with the *MusicBot*.

The critiquing settings seem to have little impact on user perception of recommendations with a conversational user interface (CUI). According to participants' responses to the post-study questionnaires (see Figure 5), we saw that they have similar perceptions of recommendations in HC and UC, except *the effort of looking for songs*. Surprisingly, we find users felt they needed more effort to look for a song using HC than using UC.

The interaction results show that HC tends to lead to more completion time, more listened songs, and more dialog turns, which might indicate that HC may increase user engagement. Moreover, the in-depth analysis of the role of SC implies that users tend to perceive higher diversity and ease of use if they have tried SC in the condition of HC. The higher conversation rate of SC also implies that SC is more effective than UC in finding liked songs.

Overall, the subjective responses suggest that both UC and HC are useful for users to find good quality songs, but HC may increase user engagement and possibility to find diverse music. Thus, we answer the research question **RQ1**: *Which critiquing setting, UC versus HC, is better suited for controlling music recommendations?*

***Our findings suggest that combining UC and SC in a conversational user interface may increase user engagement and likelihood of finding more (diverse) songs***.

Both musical sophistication (MS) and desire for control (DFC) positively influence user perception of recommendations in a CUI. Previous studies have shown that MS positively affects perceived quality, which in turn leads to a higher recommendation acceptance [20]. Likewise, our results also suggest users with higher MS tend to find more songs matching their interests. In addition, with higher MS, users are more likely to feel in control, provide feedback to recommendations, trust the recommended items, and reuse the system, which have not been found in previous studies conducted with GUIs. Arguably, CUIs may stimulate users with high MS to make critiques and to look for better recommendations they would like to trust. Previous studies [17, 22] have shown that systems implementing user control can increase user perceived understanding, their rating of the recommendation, and satisfaction. However, to the best of our knowledge, little work investigated how DFC influences the perceived quality. Overall, users with high DFC are more likely to feel in control of modifying the system, and they think *MusicBot* is easy to understand and know how to communicate with the bot. However, it seems that PC has little impact on user interaction behavior. The only significant correlation indicates that higher MS tends to lead to higher ratings, which is somewhat in line with the reported positive effect of MS on acceptance [20]. Although we also found users with high tech saviness (TS) felt they need more effort to find a good song, we do not clearly know an explanation for this correlation. We argue that users with high TS may have higher expectation of the conversational agent. However, a recent study [25] shows the Gulf between user expectation and experience with conversational agents, which may influence users to gauge the effort they spend on looking for a song. Thus, we have answered the research question **RQ2**: *Which personal characteristics might influence the user's perception and interaction of recommendations?*

***We suggest system designers should consider MS and DFC as key personal characteristics that may influence the conversational interaction design for critiquing-based music recommendations.***

Most of existing critiquing-based recommender systems have been designed for making critiques for high-value products such as computers and digital cameras, with the purpose of avoiding users' financial risk. Our study investigates whether the critiquing element could also be useful in the low-involvement product domain such as music. Pu et al. [32] demonstrate a system that combines critiques, public opinions, and expert advice to improve user decision confidence for low-involvement products such as perfume. In our study, users' overall responses to the post-study questionnaire suggest that no matter the critiquing setting, our *MusicBot* were perceived at a "good level" (above 4.5 on a 7-point Likert scale) for most UX metrics. Furthermore, the user subjective feedback in Section 5.4 reflects users' positive attitude towards employing critiquing techniques to search for music with *MusicBot*. We have answered the research question **RQ3**: *Are critiquing techniques perceived as useful in low-involvement product domains as in high-involvement product domains?*

***We suggest system designers to incorporate a proper critiquing technique for low-involvement domains to augment recommendations such as for music and movies.***

## 7 LIMITATIONS

First, we provided a public Spotify account for the sake of some participants who are not active Spotify users. For these users, instead of retrieving the user profile from a Spotify account, we asked users to build their user profiles manually, which may be biased by their engagement in the study. Ultimately, the quality of user profile may affect the actual quality of recommendations.

Second, we predefined seven user intents for the supported critiques and music play control in Dialogflow. However, users who expect *MusicBot* to have more capability may perceive it not "smart"

enough to understand their intention, but this is beyond the current scope of *MusicBot*.

Third, the sample size in our study is relatively small, which may undermine the power of the statistical analysis.

## 8 CONCLUSION

In this paper, we presented an online evaluation of two different critiquing settings (UC and HC) implemented in a conversational agent for music recommendations. Generally speaking, the recommendations generated by UC and HC were perceived equally by users in terms of several UX metrics. However, compared with UC, HC tends to increase user engagement in searching for a song, which might be attributed to more dialog turns, listened songs, and completion time found in the HC system. Moreover, we found two personal characteristics, music sophistication (MS) and desire for control (DFC), which positively influence user perceptions of recommendations. For user interaction, we only found one positive correlation between MS and user ratings. Finally, our study validated the usefulness of critiquing based conversational systems for low-involvement product domains.

Overall, relative to existing research on critiquing-based recommender systems (CBRS) with graphical user interfaces, our research attempts to evaluate various critiquing techniques with conversational interaction. In the future, we plan to investigate other possible personal characteristics that may influence the way users interact with recommendations in such a conversational agent.

## REFERENCES


[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proc. of SIGMOD'93*, Vol. 22. ACM, 207–216.

[2] Timothy W Bickmore, Dina Utami, Robin Matsuyama, et al. 2016. Improving access to online health information with conversational agents: a randomized controlled experiment. *Journal of medical Internet research* 18, 1 (2016), e1.

[3] Jerry M Burger. 1986. Desire for control and the illusion of control: The effects of familiarity and sequence of outcomes. *Journal of research in personality* 20, 1 (1986), 66–76.

[4] Jerry M Burger and Harris M Cooper. 1979. The desirability of control. *Motivation and emotion* 3, 4 (1979), 381–393.

[5] Robin Burke. 2000. Knowledge-based recommender systems. *Encyclopedia of library and information systems* 69, Supplement 32 (2000), 175–186.

[6] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *UMUAI* 12, 4 (2002), 331–370.

[7] Li Chen and Pearl Pu. 2007. Hybrid critiquing-based recommender systems. In *Proc. of IUI'07*. ACM, 22–31.

[8] Li Chen and Pearl Pu. 2007. Preference-based organization interfaces: aiding user critiques in recommender systems. In *Proc. of UM'07*. Springer, 77–86.

[9] Li Chen and Pearl Pu. 2009. Interaction design guidelines on critiquing-based recommender systems. *UMUAI* 19, 3 (2009), 167.

[10] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. *UMUAI* 22, 1-2 (2012), 125–150.

[11] Eric Corbett and Astrid Weber. 2016. What can I say?: Addressing user experience challenges of a mobile voice user interface for accessibility. In *Proc. of MobileHCI'16*. ACM, 72–82.

[12] Lei Cui, Shaohan Huang, Furu Wei, et al. 2017. Superagent: A customer service chatbot for e-commerce websites. *Proc. of ACL'17, System Demonstrations* (2017), 97–102.

[13] Hendrik Drachsler, Toine Bogers, Riina Vuorikari, et al. 2010. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science* 1, 2 (2010), 2849–2858.

[14] Ahmed Fadhil. 2018. Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation. *arXiv preprint arXiv:1802.09100* (2018).

[15] Peter Grasch, Alexander Felfernig, and Florian Reinfrank. 2013. Recomment: Towards critiquing-based recommendation with speech interaction. In *Proc. of RecSys'13*. ACM, 157–164.

[16] David B Greenberger, Stephen Strasser, and Soonmook Lee. 1988. Personal control as a mediator between perceptions of supervisory behaviors and employee reactions. *Academy of Management Journal* 31, 2 (1988), 405–417.

[17] F Maxwell Harper, Funing Xu, Harmanpreet Kaur, et al. 2015. Putting users in control of their recommendations. In *Proc. of RecSys'15*. ACM, 3–10.

[18] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *EC-Web'16*. Springer, 21–33.

[19] Yucheng Jin, Karsten Seipp, Erik Duval, et al. 2016. Go with the flow: Effects of transparency and user control on targeted advertising using flow charts. In *Proc. of AVI'16*. ACM, 68–75.

[20] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proc. of RecSys'18*. ACM, 13–21.

[21] Jie Kang, Kyle Condiff, Shuo Chang, et al. 2017. Understanding how people use natural language to ask for recommendations. In *Proc. of RecSys'17*. ACM, 229–237.

[22] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, et al. 2012. Inspectability and control in social recommenders. In *Proc. of RecSys'12*. ACM, 43–50.

[23] Tobias Kowatsch, Marcia Nißen, Chen-Hsuan I Shih, et al. 2017. Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity. In *Proc. of PEACH'17*. ETH Zurich.

[24] Greg Linden, Steve Hanks, and Neal Lesh. 1997. Interactive assessment of user preference models: The automated travel assistant. In *Proc. of UM'97*. Springer, 67–78.

[25] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proc. of CHI'16*. ACM, 5286–5297.

[26] Kevin McCarthy, James Reilly, Lorraine McGinty, et al. 2004. On the dynamic generation of compound critiques in conversational recommender systems. In *Proc. of AH'04*. Springer, 176–184.

[27] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, et al. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proc. of IUI'19*. ACM, 397–407.

[28] Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, et al. 2018. Controlling Spotify recommendations: Effects of personal characteristics on music recommender user Interfaces. In *Proc. of UMAP'18*. ACM, 101–109.

[29] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. 2014. The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one* 9, 2 (2014), e89642.

[30] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proc. of RecSys'11*. ACM, 157–164.

[31] Pearl Pu, Li Chen, and Pratyush Kumar. 2008. Evaluating product search and recommender systems for E-commerce environments. *Electronic Commerce Research* 8, 1-2 (2008), 1–27.

[32] Pearl Pu, Maoan Zhou, and Sylvain Castagnos. 2009. Critiquing recommenders for public taste products. In *Proc. of RecSys'09*. ACM, 249–252.

[33] James Reilly, Kevin McCarthy, Lorraine McGinty, et al. 2004. Dynamic critiquing. In *European Conference on Case-Based Reasoning*. Springer, 763–777.

[34] James Reilly, Kevin McCarthy, Lorraine McGinty, et al. 2004. Incremental critiquing. In *Proc. of SGAI'04*. Springer, 101–114.

[35] Daniel Schulman and Timothy Bickmore. 2009. Persuading users through counseling dialogue with a conversational agent. In *Proc. of Persuasive Technology'09*. ACM, 25.

[36] Hideo Shimazu. 2001. ExpertClerk: Navigating shoppers' buying process with the combination of asking and proposing. In *Proc. of IJCAI'01*. Morgan Kaufmann Publishers Inc., 1443–1448.

[37] Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research* 21 (2004), 393–428.

[38] Paolo Viappiani, Boi Faltings, and Pearl Pu. 2006. Evaluating preference-based search tools: a tale of two approaches. In *Proc. of AAAI'06*. AAAI press, 205–211.

[39] Marilyn A Walker, Diane J Litman, Candace A Kamm, et al. 1997. PARADISE: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004* (1997).

[40] Anbang Xu, Zhe Liu, Yufan Guo, et al. 2017. A new chatbot for customer service on social media. In *Proc. of CHI'17*. ACM, 3506–3510.

[41] Longqi Yang, Michael Sobolev, Christina Tsangouri, et al. 2018. Understanding user interactions with podcast recommendations delivered via voice. In *Proc. of RecSys'18*. ACM, 190–194.

[42] Jiyong Zhang and Pearl Pu. 2006. A comparative study of compound critique generation in conversational recommender systems. In *Proc. of AH'06*. Springer, 234–243.

[43] Chunyi Zhou, Yuanyuan Jin, Kai Zhang, et al. 2018. MusicRoBot: Towards conversational context-aware music recommender system. In *Proc. of DASFAA'18*. Springer, 817–820.