# Critiquing for Music Exploration in Conversational Recommender Systems

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

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

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

## ABSTRACT

Dialogue-based conversational recommender systems allow users to give language-based feedback on the recommended item, which has great potential for supporting users to explore the space of recommendations through conversation. In this work, we consider incorporating critiquing techniques into conversational systems to facilitate users' exploration of music recommendations. Thus, we have developed a music chatbot with three system variants, which are respectively featured with three different critiquing techniques, i.e., *user-initiated critiquing (UC)*, *progressive system-suggested critiquing (Progressive SC)*, and *cascading system-suggested critiquing (Cascading SC)*. We conducted a between-subject study (N=107) to compare these three types of systems with regards to music exploration in terms of user perception and user interaction. Results show that both UC and SC are useful for music exploration, while users perceive higher diversity of recommendations with the system that offers *Cascading SC* and perceive more serendipitous with the system that offers *Progressive SC*. In addition, we find that the critiquing techniques significantly moderate the relationships between some interaction metrics (e.g., number of listened songs, number of dialogue turns) and users' perceived helpfulness and serendipity during music exploration.

## CCS CONCEPTS

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

## KEYWORDS

conversational recommender systems; music exploration; critiquing; conversational interaction

**ACM Reference Format:**
Wanling Cai, Yucheng Jin, and Li Chen. 2021. Critiquing for Music Exploration in Conversational Recommender Systems. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450657

*IUI '21, April 14–17, 2021, College Station, TX, USA*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450657

## 1 INTRODUCTION

Recommender systems have become critically important for helping users quickly find ideal items among a large number of products [25]. However, personalized recommendations may lead users to increasingly narrower space of items over time (called "filter-bubble" effects) [20, 30]. To mitigate this issue, several attempts have been made to encourage users to explore diverse sets of items, such as diversity-driven algorithms [41, 42] and visualizing recommendations [16, 32]. On the other hand, dialogue-based conversational recommender systems enable users to freely give feedback on recommendations through natural language [3, 11, 12], which show considerable potential for promoting users' exploratory activities. However, so far little work has studied supporting user exploration through conversational interaction.

Recently, there is a work [12] that studied a chatbot for accommodating user control in music recommendations with two critiquing techniques (i.e., ***user-initiated critiquing (UC)*** and ***system-suggested critiquing (SC)*** [7]). The user study of this work reveals that users tend to feel receiving more diverse recommendations when using the system with both UC and SC. Inspired by this observation, we consider to stimulate users' exploration of recommendations by strengthening the critiquing technique in conversational interaction.

Therefore, in the current work, we have designed two kinds of system-suggested critiquing technique: ***Progressive system-suggested critiquing (Progressive SC)*** and ***cascading system-suggested critiquing (Cascading SC)*** for facilitating users' exploration of music with two different directions: The former is preference-oriented, which provides critiques based on users' current preferences and incremental critiquing feedback [24], while the latter is diversity-oriented, which suggests critiques to steer users into a cascade of diverse types of music using a strategical approach with the assumption of the cascading user behavior as inspired by [19]. Then, we have developed a music chatbot with three system variants, which are respectively featured with ***UC*** (i.e., users can make critiques on the recommended songs to explore songs they want), ***Progressive SC*** and ***Cascading SC***. To investigate how these critiquing techniques influence users' music exploration with conversational interaction, we conducted a between-subject user study (involving 107 participants) to compare the three system variants in terms of both user perception of and user interaction with recommendations. We also examined how these critiquing techniques moderate the relationship between user interaction behavior and user perception of music recommendations.

In a short summary, we have mainly focused on answering two research questions as follows (see Figure 1):

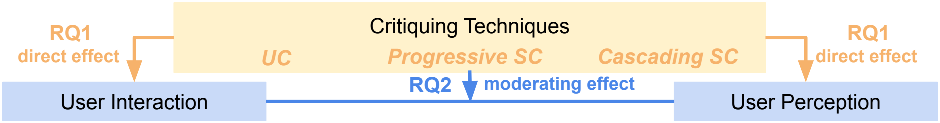


**Figure 1: Our research questions.**

**RQ1:** *How do critiquing techniques influence users' exploration of music in a conversational recommender?*

**RQ2:** *How do critiquing techniques moderate the relationship between user interaction behavior and user perception of music recommendations?*

Our main contributions of this work are four-fold:

(1) We have proposed two kinds of system-suggested critiquing technique, in order to encourage users' exploration of music recommendations, and compared three variants of the system supported with different critiquing methods (i.e., *UC*, *Progressive SC*, and *Cascading SC*) in terms of users' perception of and interaction with recommendations. The experimental results show that users perceive higher diversity of recommendations with the system that offers *Cascading SC* and feel more serendipitous recommendations with the system that offers *Progressive SC*.

(2) We have investigated the moderation effects of critiquing techniques, and find that the critiquing techniques significantly moderate some relationships between interaction metrics (such as number of listened songs and number of dialogue turns) and user perception metrics (such as perceived helpfulness and serendipity).

(3) We have analyzed users' interaction flow towards UC and SC, and find that users tend to use UC when they have gradually established their new preferences during the interaction with conversational recommendations, while users may be stimulated to request SC when they have benefited from the SC proactively offered by the system.

(4) We have discussed our findings and provided practical implications for designing a critiquing-based conversational recommender system for supporting users' music exploration.

# 2 RELATED WORK

## 2.1 User Exploration in Recommender Systems

Prior work has shown various strategies to support user exploration by diversity-driven algorithms or visualizing recommendations. Diversity-driven algorithms typically generate recommendations that maintain the balance between accuracy and diversity [9, 18, 40–42]. For instance, some researchers proposed to increase the recommendation diversity based on items' attributes, such as book topics [42], movie genres, and social tags [33]. In [30], the authors proposed a way to help users take a gradual path towards the desired new music preference by traversing user preference graphs and generating a sequence of artists as guided transition. Most of the related studies have attempted to increase the diversity for a ranked list, but there are some limitations [6, 29], such that users tend to pay less attention to the bottom of the list when exploring recommendations, which is a position bias. Besides, some works have visualized recommendations to support user exploration in recommender systems. For example, to raise users' awareness of exploration, [16, 32] highlighted the regions of the underrepresented recommendation space, so-called blind spots, which could help users to identify what is known and what is unknown in their profile. Some visualization systems, such as TalkExplorer, [34], and Moodplay [1], allow users to explore diverse items during the recommendation process. In [26, 27], the authors introduced a shortlist as a short-term memory to reduce users' cognitive efforts and help users make better decisions when exploring diverse movies. The recommender systems discussed so far support user exploration by presenting diverse recommendations or visualizations. To the best of our knowledge, little work has been done to support user exploration with conversational interaction.

## 2.2 Conversational Recommender Systems

Conversational recommender systems aim to help users seek for their desired items through natural language [11]. Several studies have demonstrated this kind of conversational systems [28, 31]. For instance, ExpertClerk [28] is a conversational agent designed to interact with shoppers by asking questions to obtain their preferences and proposing recommendations to assist users to find their satisfactory products. Adaptive place advisor [31] provides personalized recommendations to assist users to find preferable places for traveling by considering both users' long-term preferences and short-term interests. Also, several studies show the superiority of conversational user interfaces over graphical user interfaces during the process of recommendations [10, 15, 38].

In the broad area of recommender systems, critiquing-based recommender systems have been proposed to elicit users' critiquing feedback to help the system improve the recommendation [7]. In particular, there are two major types of critiquing technique, including user-initiated critiquing (i.e., users construct critiques by themselves) and system-suggested critiquing (i.e., the system generates a set of critique candidates for users to choose). A recent work [12] studied such kind of system with conversational interaction and found that critiquing techniques enable users to control recommendations in conversational user interfaces. They also observed that users tend to perceive higher diversity and efficiency when the conversational recommender presents system-suggested critiques compared to the system that only supports user-initiated critiquing. Inspired by this observation, we are interested in in-depth investigating how critiquing techniques can support users' exploration of recommendations with conversational interaction.

Different from [12], in this work, for stimulating users' exploratory activities, we introduce two kinds of system-suggested critiquing: *Progressive system-suggested critiquing* that is preference-oriented (generating critiques considering both users' current preferences and incremental critiquing feedback [24]); and *cascading system-suggested critiquing* that is diversity-oriented (suggesting critiques in a strategic approach with the assumption of the cascading user behavior as motivated by [19]). In addition, we consider the chatbot's proactivity in our designed systems (i.e., the ability of proactively offering SC to encourage users to explore music), since some previous studies have shown that the robot's proactivity may help people get rich information and reduce the decision space [23].

## 3 SYSTEM DESIGN

Following the workflow of an existing music chatbot [12], we have developed a music chatbot by using a popular NLU platform, DialogFlow[1], and a widely used music service, Spotify API[2]. The system supports both user-initiated critiquing (UC) and system-suggested critiquing (SC). In particular, we devise two kinds of system-suggested critiquing in the newest version: *Progressive system-suggested critiquing (**Progressive SC**)* that guides users to explore a group of songs based on their current preferences and critiquing feedback; and *Cascading system-suggested critiquing (**Cascading SC**)* that motivates users to explore a cascade of different types of music. To investigate how different critiquing techniques help users explore music recommendations, we concretely implemented three variants of the critiquing system:

**User-initiated Critiquing System (User-C)**: The system only supports UC. Users can post user-initiated critiques to actively explore songs based on music-related attributes such as genres, tempo, and danceability.

**Progressive Critiquing System (Progressive-C)**: The system is a hybrid critiquing system that supports both UC and SC. Users can either post UC or ask the system to provide *Progressive SC* to help them discover music.

**Cascading Critiquing System (Cascading-C)**: Similar to the Progressive-C system, the system also supports both UC and SC, but provides *Cascading SC* when the system-suggested critiquing is triggered.

Inspired by recent studies about the chatbot's proactivity [23], the two hybrid critiquing systems (i.e., Progressive-C and Cascading-C) are designed to provide SC in two different manners: **Reactive SC** refers to the SC that suggests critiques to users when they make an explicit request (i.e., clicking the button "Let bot suggest" during the conversation); **Proactive SC** refers to the SC that proactively offers critiques for stimulating users to explore music.

### 3.1 Behavior Policies and Algorithms

Based on the typical recommendation process introduced in [7], we design associated behavior policies for these three types of system as shown in Figure 2.

***Initiation:*** Before initiating the conversation, the system obtains users' initial preferences for three attributes, i.e., songs, artists, and music genres, so as to initialize the user model. Of note, the music data (including metadata and song attributes) in our system were obtained from the Spotify platform. Our system gets users' preference data from their profiles in Spotify or creates preference data for non-regular Spotify users by asking them the favorite songs and artists. Then, the system calls Spotify recommendation API to obtain 150 recommendations for generating a ranked playlist based on the initial user model according to the Multi-Attribute Utility Theory (MAUT) [39], which formally estimates the user (denoted as $u$)'s preference over each song (denoted as $i$) as $r^M_{u,i} = \sum_{a \in \mathcal{A}} w_{u,a} \times v(u,i,a)$, where $\mathcal{A}$ denotes all concerned music-related attributes, and $w_{u,a}$ is the relative importance (i.e., the user $u$'s preference weight) of the attribute $a$. $v(u,i,a)$ represents the user $u$'s preference over the song $i$ regarding the attribute $a$, which is measured as $p(k_{a,i}|\mathcal{I}^{liked}_u)$, i.e., the probability that the attribute $a$'s values appearing in the user $u$'s previous favorite songs ($\mathcal{I}^{liked}_u$) fall into the value bin[3] of the attribute $a$ of the currently considered song $i$ (denoted as $k_{a,i}$). The initials weights of all attributes are the same and will be gradually adjusted based on the user' subsequent critiques on the attributes.

***User-initiated Critiquing (UC):*** After receiving a recommendation, the user may make user-initiated critique on its audio attributes (i.e., energy, danceability, speechiness, tempo, and valence), music categories, or artists, e.g., saying *"I want higher tempo."* The system then updates the user model and returns a new recommendation.

***System-suggested Critiquing (SC):*** In the two hybrid critiquing systems, the user can ask for the system's suggested critiques (i.e., *Progressive SC* or *Cascading SC*) by clicking the button *"Let bot suggest"*. Then, the system provides the suggested critique for assisting the user to explore music recommendations, e.g., *"Compared to the last played song, do you like the song of lower tempo?"* User feedback to the suggested critique (*Accept* by clicking the button "Yes" or *Reject* via the button "No") will be used to update the user model and make subsequent recommendations.

There are two major differences between ***Progressive SC*** and ***Cascading SC***. First, the critique selection of *Progressive SC* mainly considers the user's preference over songs and critiquing feedback as captured from the previous interactions, while *Cascading SC* focuses more on the diversity of recommended songs. Second, *Cascading SC* contains two levels of critiquing for exploring diverse songs: At Level 1, the suggested critiques are on audio features, which keep the user exploring songs within the current music genre; at Level 2, critiques are on music genres, which encourage the user to try songs in a different genre. *Progressive SC*, however, does not make a distinction between audio attributes and genres.

Specifically, the generation of these two kinds of system-suggested critique consists of the following four steps:

(1) The system first constructs a critique pattern vector for each candidate song in the current playlist (e.g., *{(genre, pop), (valence, higher), ..., (danceability, lower)}*) by comparing it with the currently recommended song in terms of music-related attributes. Each critique pattern (e.g., *(genre, pop)*) denotes a critique that contains one attribute, which is also called unit critique [7].

(2) The system filters out the critiques rejected by the user in her/his previous interactions, as well as the critiques rarely occurring in all critique pattern vectors (frequency lower than 10%). Then, for each remaining critique, the songs in the current playlist that satisfy this critique are grouped together as its contained songs.

(3) The system selects ***Progressive SC*** by calculating the utility of each remaining critique (denoted as $c$) [5] as $U_u(c) = w_{u,a_c} \times f_c \times \frac{1}{\mathcal{I}_c} \sum_{i \in \mathcal{I}_c} (r^M_{u,i} + r^C_{u,i})$, where $w_{u,a_c}$ denotes

[1] https://dialogflow.com

[2] https://developer.spotify.com/documentation/web-api

[3] We divided the value range of each attribute into 10 or 15 bins for numerical attributes. For categorical attributes, each value refers to one value bin.

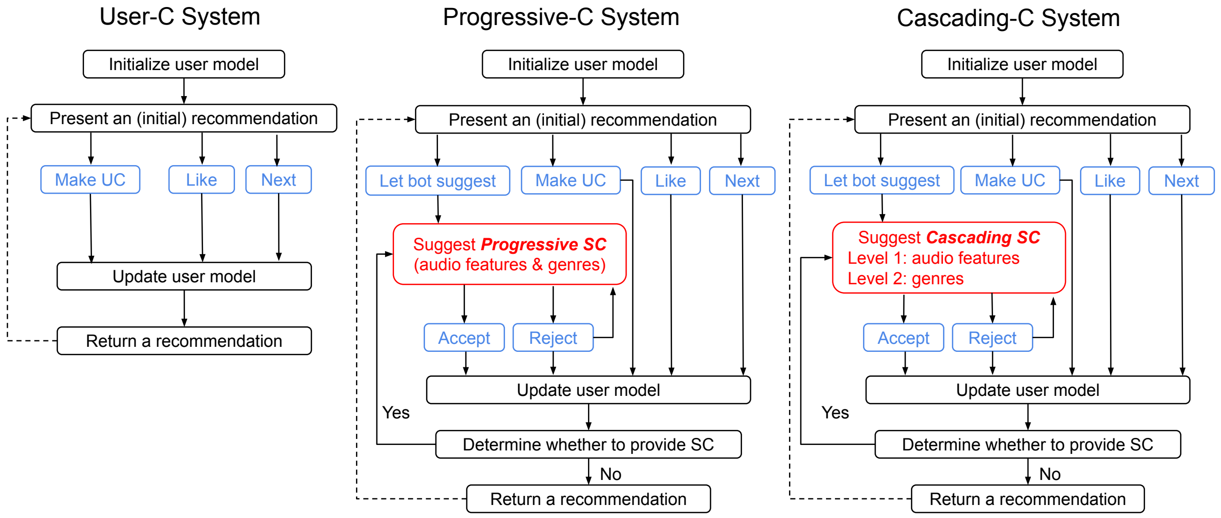


**Figure 2: Three system variants' behavior policies during the recommendation process.**

the user $u$' preference for $c$'s contained attribute, $f_c$ denotes the relative frequency of $c$ among all critique pattern vectors, and $\mathcal{I}_c$ denotes the set of songs that satisfy $c$. $\frac{1}{\mathcal{I}_c}\sum_{i\in\mathcal{I}_c}(r^M_{u,i}+r^C_{u,i})$ represents $u$'s preference over $c$'s contained songs, which considers $u$'s preference over the song $i$ (estimated as $r^M_{u,i}$ based on MAUT), as well as the compatibility of $i$ with the critiques previously made by $u$ ($PC_u$) [24] (calculated as $r^C_{u,i} = \frac{1}{|PC_u|}\sum_{c'\in PC_u} satisfies(c', i)$, where $satisfies(c', i)$ is an indicator function used to check whether the song $i$ satisfies $c'$). For ***Cascading SC***, the system calculates the overall diversity of the critique's contained songs and the songs the user has listened to in the previous interactions, for which the diversity is calculated by the average Shannon's entropy across all music-related attributes [37]: $D_u(c) = \sum_{a\in\mathcal{A}} H_a(c)$, where $H_a(c) = -\sum_{k\in K_a} p(k|\mathcal{I}_c \cup \mathcal{I}_L)\log p(k|\mathcal{I}_c \cup \mathcal{I}_L)$ measures the entropy[4] of the attribute $a$, $k \in K_a$ denotes one value bin $k$ in all value bins $K_a$ of the attribute $a$, $\mathcal{I}_L$ denotes the listened songs by the user, $\mathcal{I}_c \cup \mathcal{I}_L$ represents the resulting set of songs when the user accepts $c$, and $p(k|\mathcal{I}_c \cup \mathcal{I}_L)$ refers to the probability that the attribute $a$'s values of the resulting set of songs fall into the value bin $k$. Motivated by observations of our pilot study[5], we determine *Cascading SC* will be switched from Level 1 to Level 2 when the user likes more than 4 songs or skips more than 3 songs within the currently explored music genre.

(4) The system finally shows the critique of the highest utility $U(c)$ in Progressive-C or diversity $D(c)$ in Cascading-C.

***User Modeling.*** User model contains two parts: (1) **user preference model** stores the user's preferred value range and preference weight for the critiqued attribute, i.e., a music genre or an audio feature, which will be adjusted based on the user's feedback on the recommended item (i.e., clicking *"Like"* for accepting or *"Next"* for skipping) and the critique made by the user; (2) **user critiquing history** tracks all occurred critiques in the current dialogue.

***Dialogue Management.*** All the three systems are designed to respond to the user's inputs after detecting her/his intents, but they may respond differently to the detected intents. For the User-C system, the system proceeds to the next recommendation based on the user's intent, while the two hybrid systems will determine whether it is time to recommend a song or show a system-suggested critique based on the user's interaction behavior. We find it is reasonable to let the system proactively offer critique if the user has consecutively skipped 3 recommended songs or listened to 5 songs according to our observations in the pilot study.

***Recommendation.*** With the refined user model, we re-rank the current playlist by considering the estimated user preference over each candidate song based on MAUT and each song's compatibility with the user's critiquing feedback.

### 3.2 User Interface Design

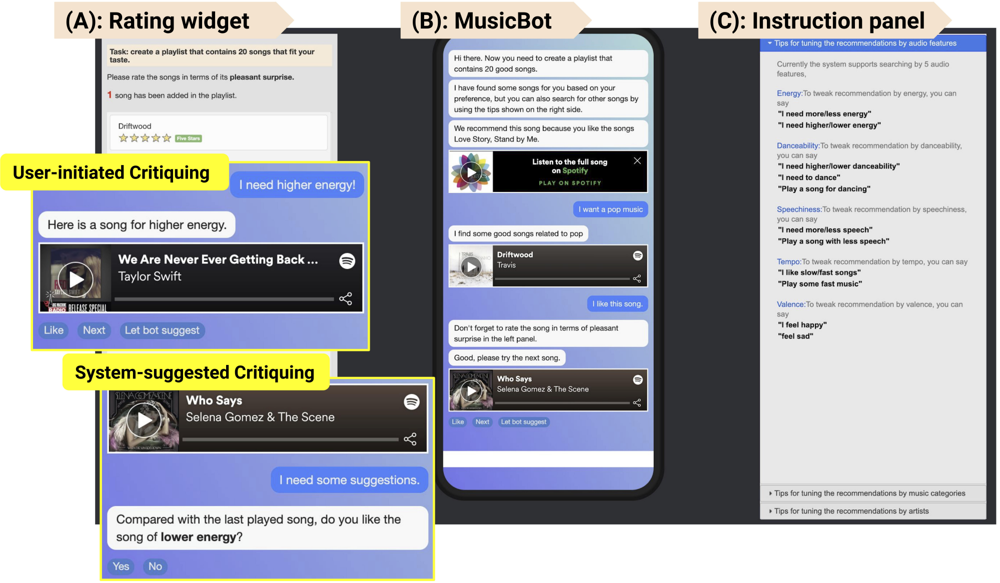
The user interface of the music chatbot consists of three parts: a rating widget, a dialogue window, and an instruction panel. Specifically, the dialogue window (Figure 3, B) shows the dialogue between the user and the bot. The recommended song is shown on a card with a set of buttons under the card for the user to give feedback. When the user clicks the "Like" button, the current song will be added to the playlist where the user can rate the song (Figure 3, A). The "Next" button allows the user to skip the current song, and the "Let bot suggest" button is to activate a system-suggested critique on the currently recommended song. If the user would like to critique the recommended song on her/him own, s/he can send a message to tune the recommendation by audio features, music categories, or artists (Figure 3, C explains the supported features with some examples). Two dialogue examples illustrate how the user can make user-initiated critiquing (UC) and system-suggested critiquing (SC) respectively.

## 4 EVALUATION

To investigate how different types of critiquing technique influence user exploration of music recommendations, we created three experimental conditions (respectively corresponding to User-C, Progressive-C, and Cascading-C) and conducted an online user

[4] A higher entropy of an attribute indicates that the resulting set contains songs with higher diversity in terms of this attribute.

[5] We conducted a lab controlled pilot study (with 3 volunteers) in order to test adequacy of our system and the experimental procedure.



**Task: create a playlist that contains 20 songs that fit your taste.**

Please rate the songs in terms of its **pleasant surprise.**

1 song has been added in the playlist.

Driftwood ☆☆☆☆☆

I need higher energy!

Here is a song for higher energy.

We Are Never Ever Getting Back ... Taylor Swift

Who Says Selena Gomez & The Scene

I need some suggestions.

Compared with the last played song, do you like the song of **lower energy**?

Hi there. Now you need to create a playlist that contains 20 good songs.

I have found some songs for you based on your preference, but you can also search for other songs by using the tips shown on the right side.

We recommend this song because you like the songs Love Story, Stand by Me.

I want a pop music

I find some good songs related to pop

Driftwood Travis

I like this song.

Don't forget to rate the song in terms of pleasant surprise in the left panel.

Good, please try the next song.

Who Says Selena Gomez & The Scene

Tips for tuning the recommendations by audio features

Currently the system supports searching by 5 audio features,

Energy:To tweak recommendation by energy, you can say
**"I need more/less energy"**
**"I need higher/lower energy"**

Danceability:To tweak recommendation by danceability, you can say
**"I need higher/lower danceability"**
**"I need to dance"**
**"Play a song for dancing"**

Speechiness:To tweak recommendation by speechiness, you can say
**"I need more/less speech"**
**"Play a song with less speech"**

Tempo:To tweak recommendation by tempo, you can say
**"I like slow/fast songs"**
**"Play some fast music"**

Valence:To tweak recommendation by valence, you can say
**"I feel happy"**
**"feel sad"**

Tips for tuning the recommendations by music categories

Tips for tuning the recommendations by artists

**Figure 3: The user interface of our music chatbot. Note that the user interface is the same as that of [12], but the underlying algorithms used to generate the two kinds of system-suggested critique (*Progressive SC* and *Cascading SC*) are different.**

study (N=107) with a between-subjects design. We randomly assigned participants to one of the three experimental conditions. The numbers of participants in the three conditions are 35, 36, and 36, respectively.

## 4.1 Participants

Participants were recruited from the Prolific platform[6], which is popularly used for academic surveys [22]. To ensure the quality of the experiment, we pre-screened users in Prolific using the following criteria: (1) participants should be fluent in English; (2) number of her/his previous submissions should be more than 100; (3) approval rate should be greater than 95%. The experiment took 25 mins on average and each participant was compensated £2.4 if s/her successfully completed the experiment.

A total of 147 users participated in our study, which is within our estimated sample size[7]. 22 participants' responses were removed since their data were detected as outliers for extremely long duration, and 18 participants were filtered out due to failure to pass the attention check questions. We finally kept the data of 107 participants (Gender: Female = 52, Male = 53, Other: 2; Age: 19-25(40), 26-30(19), 31-35(16), 36-40(9), 41-50(13), 51-60(8), > 60(2)). Participants are from different countries, including United Kingdom (35), Portugal (16), United States of America (12), Poland (11), Italy (9), Spain (4), and others (20) (e.g., Greece, Estonia, and Germany).

## 4.2 Procedure

First, participants need to accept General Data Protection Regulation (GDPR) consent form before signing into our system with their Spotify accounts. After reading the instructions of the user study, participants are asked to fill out a pre-study questionnaire. To ensure that participants understand the study task and the use of our chatbot, they read a tutorial of interacting with music recommendations in the chatbot and then try the bot for two minutes. Once they are ready, they are asked to complete the **experimental task** which contains two steps: (1) Please use our MusicBot to discover songs in different music types as much as possible, and create a playlist that contains 20 pieces of music that fit your taste, and then rate each song in terms of its pleasant surprise. (2) Then, please select top-5 most preferred songs from the created playlist. After finishing the task, participants fill out a post-study questionnaire regarding their experience with the music chatbot (see Section 4.3).

## 4.3 Measurement

The post-study questionnaire contains 10 statements (see Table 1) that measure user perception of music recommendations when using the chatbot: Q1-Q6 and Q9 are adapted from ResQue (a widely used user-centric evaluation framework for recommender systems) [4]. The statements of Q8 and Q10 [21] measure user perceived serendipity and Q7 [36] measures user engagement. Besides, the questionnaire includes three open-ended questions about music exploration: *“When do you, or why do you want to discover new songs when listening to music?” “What do you think is the difference between using the chatbot and your previous methods for discovering songs?” “Do you think whether the music chatbot could help you discover new and unexpected but interesting songs? And explain how it helps?”*

# 5 ANALYSIS & RESULTS

## 5.1 User Perception

We analyzed users’ responses to the ten statements (see Table 1) respectively in the three experimental conditions. Since the results of the Shapiro-Wilk test show that the data are not normally distributed, we performed the non-parametric one-way ANOVA

[6] https://www.prolific.co/

[7] Based on the results from our online pilot study (performed on Prolific before our main study that involves 20 participants), we calculated the sample size as 111-159 in a priori power analysis for an ANOVA F test (given a significance level $\alpha$ = .05, a power level (1-$\beta$) = .8 and an expected effect size $f$ = .25 or .3) using G*Power [8].

**Table 1: Post-study questionnaire for measuring users' perception of the music chatbot**

| Metric | Statement (each is rated on a 7-point Likert scale) |
|---|---|
| **Interest** | Q1. The songs recommended to me matched my interests. |
| **Novelty** | Q2. The songs recommended to me are novel. |
| **Music discovery** | Q3. The music chatbot helped me discover new songs. |
| **Diversity** | Q4. The songs recommended to me are diverse. |
| **Control** | Q5. I feel in control of modifying my taste using this music chatbot. |
| **Helpfulness** | Q6. The music chatbot gave me good suggestions for helping me discover songs. |
| **Engagement** | Q7. I feel it is entertaining and interesting to engage in a dialogue with this music chatbot to discover songs. |
| **Serendipity** | Q8. The music chatbot provided me with recommendations that I had not considered in the first place but turned out to be a positive and surprising discovery. |
| **Confidence** | Q9. I am confident that I will like the songs in the created playlist (20 songs). |
| **Pleasant surprise** | Q10. The songs in the created playlist (20 songs) are overall pleasantly surprising to me. |

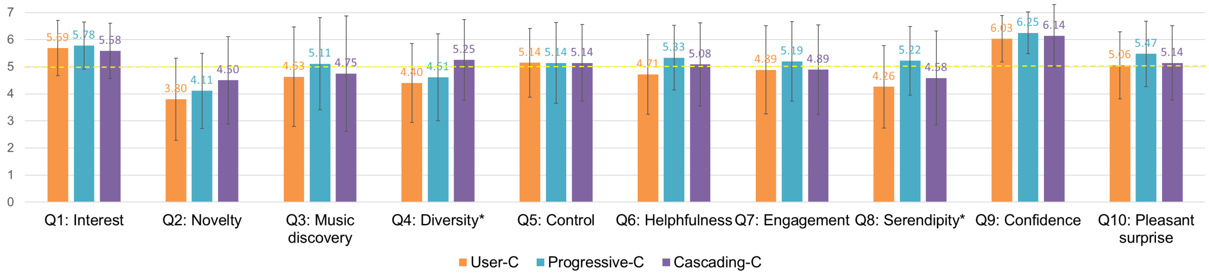


**Figure 4: Assessment results of statements related to user perception. A cut off value at 5 represents agreement on the 7-point Likert scale. * is marked for significant difference at the 5% level ($p$-value < 0.05).**

Kruskal-Wallis test for comparative analysis. It mainly indicates that the differences in terms of perceived diversity ($H$=6.81, $df$=2, $p$<.05) and perceived serendipity ($H$=7.64, $df$=2, $p$<.05) are significant among the three conditions. The post-hoc Mann-Whitney tests with Bonferroni corrected p-value show that users perceive more diversity of recommendations in *Cascading-C* ($M$=5.25, $SD$=1.48) than in *User-C* ($M$=4.40, $SD$=1.46, $p$<.05), and more serendipity in *Progressive-C* ($M$=5.22, $SD$=1.27) than in *User-C* ($M$=4.26, $SD$=1.52, $p$=.01), but no significance is found in other pairwise comparisons. This may be explained by that *Progressive SC* in *Progressive-C* can bring users different songs that are close to their interests, while *Cascading SC* in *Cascading-C* aims to introduce new types of music to users.

For the non-significant results reported in Figure 4, we still find that users positively rated all of the three system variants in some metrics related to music exploration with average ratings above 5 on the 7-point Likert scale, such as interest matching, control, confidence, and pleasant surprise. Users' perceived novelty of recommendations is relatively low in *User-C*, probably because, compared to it that only supports user-initiated critiquing, *Progressive-C* and *Cascading-C* might introduce users to more new songs with system-suggested critiques.

## 5.2 User Interaction

*5.2.1 Interaction Metrics.* We analyzed participants' interaction behavior to examine how users interacted with the three critiquing systems for music exploration. We extracted several interaction metrics from participants' logs, and analyzed their listened songs (see Table 2). We used the same statistical methods for comparison as in Section 5.1.

The results of Kruskal-Wallis tests reveal significant differences among the three conditions in terms of dialogue turns ($H$=7.75, $df$=2, $p$<.05), times of clicking buttons ($H$=20.22, $df$=2, $p$<.001), and times of typing ($H$=6.13, $df$=2, $p$<.05). The post-hoc tests show that both *Cascading-C* and *Progressive-C* led to significantly more dialogue turns than *User-C* ($p$<.05), and users clicked significantly more buttons in *Progressive-C* and *Cascading-C* than in *User-C* ($p$<.005), probably because the design of SC may introduce more dialogue turns and button clicks.

*5.2.2 Exploration Metrics.* Table 2 summarizes the ratings of songs (in terms of **pleasant surprise**) in users' created playlists and those in their selected top-5 most preferred songs, and the number of newly explored genres in each case. It shows that participants positively rated the liked songs, with the average ratings above 4 out of 5 stars in all conditions, although there are no significant differences among *User-C*, *Progressive-C* and *Cascading-C*. Moreover, relative to users' preferred genres in their initial profiles, all the three critiquing systems allow users to explore 2 to 3 new genres as shown in their created playlists.

*5.2.3 Critiquing Behavior.* To deeply investigate the role of critiquing during music exploration, we analyzed users' interaction data with focus on their critiquing behavior. First, we analyzed the actual uses of *user-initiated critiquing (UC)* and *system-suggested critiquing (SC)* in different experimental conditions. We counted the use of SC as requested by users by clicking the "Let bot suggest" button (i.e., **Reactive SC**). Table 3 shows that participants made UC more in *User-C* than in *Progressive-C* and *Cascading-C*, and made SC more in *Cascading-C* than in *Progressive-C*. In total, we find that **95** out of 107 users made UC, and **45** out of 72 users made SC in the two hybrid conditions that provide SC.

Since SC can be triggered either by clicking the "Let bot suggest" button (Reactive SC) or being proactively suggested by the system (Proactive SC), we calculated the acceptance rates of Reactive SC and Proactive SC in both *Progressive-C* and *Cascading-C*. The results show that the acceptance rate of Reactive SC (92.62%) and the acceptance rate of Proactive SC (92.13%) in *Progressive-C* are both higher than those in *Cascading-C* (respectively 77.43% and 80.71%),

**Table 2: Descriptive statistics for user interaction behavior data (significance: $^{***}$ $p < .001$, $^{**}$ $p < .01$, $^{*}$ $p < .05$)**

| | User-C | | Progressive-C | | Cascading-C | |
|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std |
| **Interaction metrics** | | | | | | |
| #Listened songs | 42.06 | 12.92 | 39.78 | 12.97 | 41.47 | 15.62 |
| Duration (minutes) | 10.95 | 4.43 | 12.04 | 4.59 | 12.47 | 5.28 |
| #Dialogue turns (times)$^{*}$ | 43.03 | 13.86 | 52.64 | 16.44 | 54.22 | 21.30 |
| #Button (times)$^{***}$ | 33.40 | 9.65 | 46.39 | 12.69 | 47.61 | 19.08 |
| #Button-Next (times) | 13.97 | 9.40 | 12.81 | 8.52 | 13.89 | 12.22 |
| #Typing (times)$^{*}$ | 9.94 | 8.17 | 6.42 | 7.62 | 6.78 | 5.40 |
| #Words per utterance | 3.32 | 1.12 | 2.72 | 1.72 | 3.66 | 1.54 |
| **Exploration metrics** | | | | | | |
| Avg Rating (Created playlist-20) | 4.27 | 0.33 | 4.37 | 0.40 | 4.28 | 0.38 |
| Avg Rating (Top-5) | 4.70 | 0.30 | 4.72 | 0.37 | 4.74 | 0.49 |
| #NewGenres (Listened songs) | 3.83 | 2.26 | 4.19 | 2.58 | 3.97 | 2.13 |
| #NewGenres (Created playlist-20) | 2.71 | 1.62 | 2.69 | 1.45 | 3.14 | 1.88 |
| #NewGenres (Top-5) | 1.40 | 1.22 | 1.56 | 1.05 | 1.44 | 1.08 |

**Table 3: Descriptive statistics for the actual uses of UC and SC, and the provenance of liked songs in the three experimental conditions**

| | User-C | Progressive-C | Cascading-C |
|---|---|---|---|
| **Actual uses of UC** | | | |
| Percentage of making UC | 94.29% (33/35) | 75.00% (27/36) | 94.44% (34/36) |
| Average times of makng UC per user | 9.52 | 8.19 | 6.71 |
| **Actual uses of SC** | | | |
| Percentage of making SC | NA | 61.11% (22/36) | 63.89% (23/36) |
| Average times of making SC per user | NA | 2.36 | 3.18 |
| **Provenance of liked songs** | | | |
| Recommendations before critiquing | 16.03% | 7.08% | 6.94% |
| UC | 83.97% | 32.10% | 45.69% |
| Reactive SC | NA | 8.97% | 11.50% |
| Proactive SC | NA | 51.85% | 35.87% |

implying that users might be prone to accept the progressive SC that fit their current preferences [20]. Besides, the way of triggering SC seems to have little impact on user acceptance of SC.

Moreover, to investigate which kind of critique is more effective for exploring diverse songs, we analyzed the provenance of the songs liked by users (see Table 3). We find that more than half of the liked songs are from Proactive SC in *Progressive-C*, probably suggesting that users can discover favorite music through the automatically suggested progressive SC with less effort to initiate UC and SC by themselves. Therefore we see that UC is less triggered in *Progressive-C* than in the other two conditions.

To further understand when users would like to make critiques for exploring diverse music, we analyzed two major interaction flows. One flow starts from the initial recommendation until the user made the first critique (UC or SC), and the other flow is between two non-consecutive critiques made by the user. We extracted three typical interaction patterns (IPs) of using UC (IP1-IP3) and one IP of using SC (IP4). The number in the parentheses indicates the percentage of participants who followed the corresponding IP when they made UC or SC.

**IP1**: *Recommend → Like → Recommend → Like → Recommend → Make UC* (56.84%, 54/95)

**IP2**: *Recommend → Next → Recommend → Next → Recommend → Make UC* (46.32%, 44/95)

**IP3**: *System Suggest Critiques → Accept SC → Recommend → Make UC* (36.07%, 22/61)

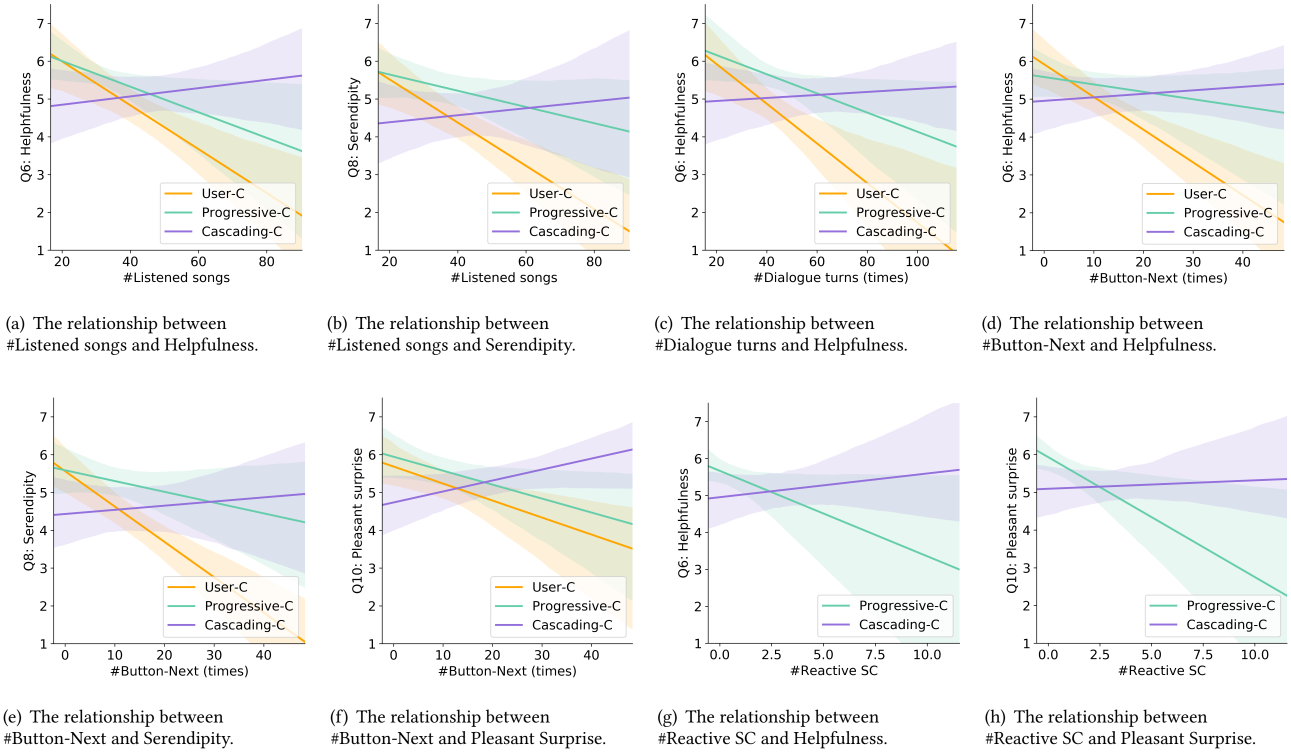
**IP4**: *System Suggest Critiques → Accept SC →...→ Let Bot Suggest* (48.89%, 22/45)

The users of IP1 and IP2 tended to use UC to explore new songs after receiving three recommended songs (that they clicked "Like" or "Next"). We identified IP3 in the two hybrid conditions *Progressive-C* and *Cascading-C* where users made UC when they felt the recommendations suggested by SC were not of their interests. The users of IP4 requested SC after accepting one or more critiques proactively suggested by the system, namely that some users are more likely to trigger reactive SC if they have benefited from proactive SC. Regarding the occurrence (times) of these IPs in the three conditions, we only observe that IP2 occurred more often in *User-C* (t=45) than in *Cascading-C* (t=24) and *Progressive-C* (t=14), probably because users can alternatively use UC or SC to adjust recommendations in the two hybrid conditions.

## 5.3 Relation Between User Interaction and User Perception

In this section, we conducted an in-depth investigation of the relationship between users' interaction behavior and their perception of the recommended songs in the three experimental conditions.

*5.3.1 Moderation Effect of Experimental Condition (EC) on the Relationship Between User Interaction and User Perception.* In order to investigate how the three experimental conditions moderate the relationship between user interaction behavior and user perception



(a) The relationship between #Listened songs and Helpfulness.

(b) The relationship between #Listened songs and Serendipity.

(c) The relationship between #Dialogue turns and Helpfulness.

(d) The relationship between #Button-Next and Helpfulness.

(e) The relationship between #Button-Next and Serendipity.

(f) The relationship between #Button-Next and Pleasant Surprise.

(g) The relationship between #Reactive SC and Helpfulness.

(h) The relationship between #Reactive SC and Pleasant Surprise.

**Figure 5: Moderation effects of experimental condition (EC) on the relationship between user interaction metrics and perception metrics.**

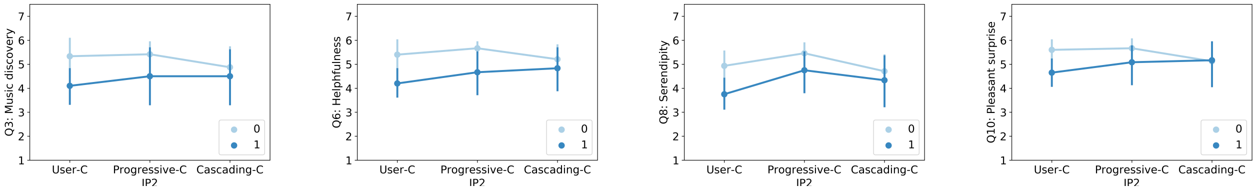
of music recommendations, we followed the two steps for moderation analysis as suggested by [35, Chapter 15]: *First*, we performed a Spearman's rank correlation analysis within each of the three experimental conditions (EC), and tested the significance of the difference between paired correlation coefficients by applying the Fisher-Z-Transformation [17]. This step serves as a preliminary analysis to assess the potential moderation of EC on the relationship between user interaction metrics and user perception metrics. *Second*, for the possible presence of moderation, we carried out a moderated regression analysis to examine the influence of EC (moderating variable) on the relationship between two variables (i.e., an interaction metric and a perception metric). Moderation effects were detected when the interaction term is statistically significant in the regression model.

As a result, EC moderates the relationships between the number of listened songs and both the perceived helpfulness ($F$(2, 101)=4.93, $p$<.01) and perceived serendipity ($F$(2, 101)=3.51 $p$<.05). Figure 5(a) and Figure 5(b) show that users who listened to more songs tended to feel the system less helpful and perceive less serendipity in the conditions *User-C* and *Progressive-C*. On the contrary, users in *Cascading-C* tended to perceive higher helpfulness and serendipity when listening to more songs. Besides, EC moderates the relationship between the number of dialogue turns and perceived helpfulness ($F$(2, 101)=4.41, $p$<.05). Figure 5(c) shows negative correlations between them in the conditions *User-C* and *Progressive-C*, but a positive correlation in the condition *Cascading-C*. Also, EC moderates the relationships between the number of "Next" button clicks and three user perception metrics: perceived helpfulness ($F$(2, 101)=4.85, $p$<.01), perceived serendipity ($F$(2, 101)=4.99, $p$<.01), and pleasant surprise ($F$(2, 101)=4.35, $p$<.05). Figures 5(d), 5(e), and 5(f) show a tendency that users who clicked more "Next" buttons seem to have lower helpfulness, serendipity and pleasant surprise in *User-C* and *Progressive-C*, while an opposite tendency is shown in *Cascading-C*. Compared with *User-C* and *Progressive-C*, *Cascading-C* can produce more diverse songs along more user interaction, which may in turn enhance user perception metrics related to music exploration. Furthermore, it shows EC moderates the relationship between the number of use of Reactive SC and both users' perceived helpfulness ($F$(1, 68)=4.30, $p$<.05) and pleasant surprise ($F$(1, 68)=6.93, $p$<.05) in *Progressive-C* and *Cascading-C*. In addition, users who more actively requested SC tended to perceive higher helpfulness and pleasant surprise in *Cascading-C* than in *Progressive-C* (see Figure 5(g) and Figure 5(h)).

In short, the above results indicate the moderation effects of EC on the relationships between some particular interaction metrics (e.g., number of listened songs, number of dialogue turns) and users' perceived helpfulness, recommendation serendipity and pleasant surprise.

### 5.3.2 Relationship Between Interaction Patterns and User Perception.

Furthermore, we investigated how the identified frequent interaction patterns (IP1-IP4) may influence user perception of recommendations. For this purpose, we split users into two groups based on the presence of a particular IP regardless of the experimental condition. We then performed a non-parametric Mann-Whitney test to compare the two groups in terms of each perception metric. The results show that the group of users who followed IP2 rated negatively than the other group regarding several perception



(a) Music Discovery.

(b) Helpfulness.

(c) Serendipity.

(d) Pleasant Surprise.

**Figure 6: The comparison of user perception between users who followed IP2 (Group 1) and users who did not follow IP2 (Group 0).**

metrics, including music discovery ($U$=1017.0, $p$<0.01), helpfulness ($U$=875.0, $p$<0.001), serendipity ($U$=1005.5, $p$<0.01), and pleasant surprise ($U$=1095.5, $p$<0.05). The in-depth analysis (see Figure 6) shows that the differences between the two groups are smaller in the two hybrid conditions (especially in *Cascading-C*) than in *User-C*, which might be because users can tune the recommendation via SC in the two conditions. Besides, the users of IP3 ($N$=22) perceived significantly higher decision confidence than the users of not following IP3 ($N$=50) in both hybrid conditions ($U$=380.0, $p$<0.05).

## 5.4 Subjective Feedback

We summarize participants' responses to the three open questions about music exploration in post-study questionnaire.

**Users' propensity towards discovering new songs**. Some participants indicated that diverse types of music bring more fun, such as *"I like to listen to playlists that someone else made to find new music because it's usually full of artists I don't know. I think it's important to branch out to different genres of music because it keeps listening interesting."* (P47, *User-C*), while some of them would like to try new music but not from different styles, e.g., *" I get stuck in playlists of similar songs. So I want to discover new music, but I don't always think of trying different styles. "* (P24, *Cascading-C*).

**Perceived differences between using a chatbot and traditional ways for discovering songs**. We identify three differences and share a few remarkable comments as follows: (1) Interacting with the chatbot could make participants feel warm and friendly, such as *"Chatbot is like speaking to a real person with suggestions and sharing emotions what you like to hear and how you like to hear with chatbot's suggestions. It's different than searching manually."* (P3, *Progressive-C*) and *"The Chatbot is more pro-active in searching for songs than I normally would be."* (P85, *Cascading-C*); (2) The chatbot might be more efficient for exploring music since it allows participants to easily indicate their preferences and then adapts to the preferences, such as *"Using this chatbot can be a more efficient and quicker way to shortlisting which songs I will give a go."* (P48, *Cascading-C*) and *"With this chatbot, I can skip the songs/artists I don't like and the chatbot will know that I don't like them so it will not suggest to me the same music/artist in the future."* (P39, *User-C*); (3) The chatbot might reduce users' efforts of exploring diverse music, such as *"Regarding the bot, It would probably be easier to find suggestions for different genres, especially those that I don't usually listen."* (P68, *Cascading-C*).

**Opinions on the used system for music exploration**. The majority of our participants ($N$=72) hold a positive attitude towards the system they used, e.g., *"Yes, because I can talk to the bot so it is easier to find the right songs".* (P22, *User-C*), *"... some other songs were completely new to me and, to my surprise, I liked them and even put some on my playlist on Spotify, this chatbot could be a good feature to have to be honest"* (P29, *Progressive-C*), and *"Yes, because it's suggestions derive from what I like and that makes me feel more open to new songs"* (P20, *Cascading-C*). However, some participants complained about the system's language recognition capability, e.g., *"It didn't recognize some of the music genres I was looking for"* (P69, *Cascading-C*).

# 6 DISCUSSIONS

In this section, we discuss our research findings based on the research questions. We also offer some practical implications for designing critiquing techniques for conversational music exploration.

**RQ1:** ***How do critiquing techniques influence users' exploration of music in a conversational recommender?*** We compared user perception of recommendations and the interaction behavior data among three experimental conditions (i.e., *User-C*, *Progressive-C* and *Cascading-C*). The incorporation of system-suggested critiquing (SC) mechanism significantly influences users' perceived diversity and perceived serendipity, both of which are key to the music exploration [14]. To be specific, *Cascading SC* is more effective in discovering diverse songs, while *Progressive SC* helps users find more songs with serendipity. Unlike the user perception metrics, the objective exploration metrics do not have a significant difference among the three critiquing techniques. The comparison results of interaction behavior data imply that SC results in more dialogue turns and button clicks, which is in line with the findings of a previous user study [12]. The more dialogue turns could mean higher user engagement [12], but which could also be subject to the design of SC as it intrinsically introduces more dialog turns and button clicks. Users tend to perform less UC when *Progressive SC* is available, as automatically prompted SC (Proactive SC) results in half of the totally liked songs in *Progressive-C*. In other words, *Progressive SC* is particularly useful for exploring music and finding 20 liked songs, which may also explain why SC produces more liked songs in *Progressive-C* than in *Cascading-C*. The way of triggering SC does not seem to influence the acceptance rate of SC. But, Proactive SC has a larger impact on music exploration than Reactive SC, probably because Reactive SC is less triggered than Proactive SC. Generally speaking, both SC and UC facilitate music exploration, but SC seems to influence user perception more than UC.

**RQ2:** ***How do critiquing techniques moderate the relationship between user interaction behavior and user perception of***

***music recommendations?*** The critiquing techniques significantly moderate the relationships between some user interaction metrics (e.g., number of listened songs, number of dialogue turns, number of "Next" button clicks) and users' perceived helpfulness, serendipity, and pleasant surprise. Arguably, in *Cascading-C*, users who interact more with the system are likely to encounter more diverse types of music and even some surprising discoveries, thereby perceiving more helpfulness and having a better exploration experience. However, in *User-C* and *Progressive-C*, more user interactions may not further increase perceived helpfulness and serendipity probably due to their exploration strategies. These strategies aim at exploring music centered around users' current preferences rather than introducing new types of music. But compared with *User-C*, *Progressive-C* weakens the negative correlation between user interaction and user perception, possibly due to the positive influence of SC on user perception. *Cascading-C* even changes the negative correlation to be positive because more user interactions can trigger new exploration beyond current user preferences.

**Interaction patterns.** The identified frequent interaction patterns (IP1-IP3) of making user-initiated critiquing (UC) suggest that users may gradually establish their new preferences during the interaction with the conversational recommendations and then have a clearer exploratory direction [7]. At the same time, IP4 implies that the perceived usefulness of proactive SC influences users' intention of making a reactive SC. The comparative analysis shows that users who have followed IP2 tend to have a negative perception of several aspects, such as music discovery, helpfulness, and serendipity, implying that the rejection of recommendations might impair user experience.

**Implications of our work.** Combining the above results, we would like to recommend the hybrid critiquing approach that incorporates both UC and SC for music exploration. UC allows users to explicitly initialize exploration when they have a clear exploration goal, while SC guides users to explore recommendations especially when they have no specific goal. Regarding the two types of SC, practitioners may choose between *Progressive SC* and *Cascading SC* according to the exploration goal, e.g., diversity-oriented exploration or serendipity-oriented exploration. Moreover, the period of exploring music may also influence the choice of SC. *Progressive SC* would be more helpful in the initial period of exploration (or short-term exploration like the task in our study) where users are probably more acceptable for the songs that are close to their current preferences. By contrast, *Cascading SC* might be more useful in the later period of exploring music (or long-term exploration). After a period of exploration, users reasonably expect to see more diverse types of song. The diversity-oriented exploration in *Cascading-C* can further diversify recommendations, in turn, positively influencing user perception.

## 7 LIMITATIONS

This study has three major limitations. First, our music chatbot only allows users to explore music through genres, artists, and audios features. However, some other attributes that have been found important to music exploration like social tags [13] and mood [2] are not considered. Second, the proactive SC in the two hybrid critiquing systems is triggered under the pre-set condition (e.g., when a user consecutively skips 3 recommended songs). However, more flexible methods to determine appropriate timing for proactively offering SC would be desired. Third, the current study has a relatively small sample size, which may undermine the power of the statistical analysis.

## 8 CONCLUSION

In conclusion, we performed an online user study to compare three types of critiquing system (i.e., *User-C*, *Progressive-C*, and *Cascading-C*) in terms of supporting users' music exploration with conversational interaction. In general, they all allow users to explore diverse songs through conversation, and system-suggested critiquing (SC) in the two hybrid systems (*Progressive-C* and *Cascading-C*) further increases perceived diversity and serendipity. The moderation analyses show that critiquing techniques exert influence on the relationship between user interaction and user perception. The identified frequent interaction patterns indicate that users tend to make critiques on their own when they have established their (new) preferences through several rounds of interaction with recommendations, but will be likely to try SC after they have benefited from the proactive SC.

Overall, compared with existing user studies on conversational recommender systems, our study has particularly focused on investigating how different critiquing techniques affect music exploration with a conversational recommender. In the future, we plan to investigate how personal characteristics such as personality affect user exploration of music when interacting with different types of critiquing system. We also intend to verify if the findings in this study can be generalized to other application domains.

## ACKNOWLEDGMENTS

This work was supported by Hong Kong Baptist University IR-CMS Project (IRCMS/19-20/D05). We also thank all participants for their time in taking part in our experiment, and reviewers for their valuable comments on our paper.

## REFERENCES

[1] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive Mood-based Music Discovery and Recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Nova Scotia, Canada) *(UMAP '16)*. ACM, New York, NY, USA, 275–279. https://doi.org/10.1145/2930238.2930280

[2] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2019. Moodplay: Interactive music recommendation based on Artists' mood similarity. *International Journal of Human-Computer Studies* 121 (2019), 142–159.

[3] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3340631.3394856

[4] Li Chen and Pearl Pu. 2006. Evaluating Critiquing-based Recommender Agents. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1* (Boston, Massachusetts) *(AAAI'06)*. AAAI Press, 157–162. http://dl.acm.org/citation.cfm?id=1597538.1597564

[5] Li Chen and Pearl Pu. 2007. Preference-based organization interfaces: aiding user critiques in recommender systems. In *International Conference on User Modeling*. Springer, 77–86.

[6] Li Chen and Pearl Pu. 2010. Eye-Tracking Study of User Behavior in Recommender Interfaces. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization* (Big Island, HI) *(UMAP'10)*. Springer-Verlag, Berlin, Heidelberg, 375–380. https://doi.org/10.1007/978-3-642-13470-8_35

[7] Li Chen and Pearl Pu. 2012. Critiquing-based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (April 2012),

125–150. https://doi.org/10.1007/s11257-011-9108-6

[8] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[9] Neil Hurley and Mi Zhang. 2011. Novelty and Diversity in Top-N Recommendation – Analysis and Evaluation. *ACM Trans. Internet Technol.* 10, 4, Article 14 (March 2011), 30 pages. https://doi.org/10.1145/1944339.1944341

[10] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

[11] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. *arXiv preprint arXiv:2004.00646* (2020).

[12] Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 951–960. https://doi.org/10.1145/3357384.3357923

[13] Mohsen Kamalzadeh, Christoph Kralj, Torsten Möller, and Michael Sedlmair. 2016. TagFlip: active mobile music discovery with social tags. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 19–30.

[14] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.

[15] 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* (Como, Italy) *(RecSys '17)*. ACM, New York, NY, USA, 229–237. https://doi.org/10.1145/3109859.3109873

[16] Jaya Kumar and Nava Tintarev. 2018. Using Visualizations to Encourage Blind-Spot Exploration. In *Recsys workshop on Interfaces and Decision Making in Recommender Systems*.

[17] Wolfgang Lenhard and Alexandra Lenhard. 2014. Hypothesis Tests for Comparing Correlations. Psychometrica. https://doi.org/10.13140/RG.2.1.2954.1367

[18] Mark Levy and Klaas Bosteels. 2010. Music recommendation and the long tail. In *1st Workshop On Music Recommendation And Discovery, ACM RecSys* (Barcelona, Spain) *(WOMRAD)*.

[19] Chang Li, Haoyun Feng, and Maarten de Rijke. 2020. Cascading Hybrid Bandits: Online Learning to Rank for Relevance and Diversity. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3383313.3412245

[20] Yu Liang and Martijn C. Willemsen. 2019. Personalized Recommendations for Music Genre Exploration. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. ACM, New York, NY, USA, 276–284. https://doi.org/10.1145/3320435.3320455

[21] Christian Matt, Alexander Benlian, Thomas Hess, and Christian Weiß. 2014. Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations. In *Thirty Fifth International Conference on Information Systems*.

[22] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[23] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. 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, 1–13. https://doi.org/10.1145/3290605.3300328

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

[25] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2015. *Recommender Systems Handbook* (2nd ed.). Springer-Verlag.

[26] Tobias Schnabel, Paul N. Bennett, Susan T. Dumais, and Thorsten Joachims. 2016. Using Shortlists to Support Decision Making and Improve Recommender System Performance. In *Proceedings of the 25th International Conference on World Wide Web* (Montr&#233;al, Qu&#233;bec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 987–997. https://doi.org/10.1145/2872427.2883012

[27] Tobias Schnabel, Paul N. Bennett, Susan T. Dumais, and Thorsten Joachims. 2018. Short-Term Satisfaction and Long-Term Coverage: Understanding How Users Tolerate Algorithmic Exploration. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. ACM, New York, NY, USA, 513–521. https://doi.org/10.1145/3159652.3159700

[28] Hideo Shimazu. 2001. ExpertClerk: Navigating Shoppers' Buying Process with the Combination of Asking and Proposing. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (Seattle, WA, USA) *(IJCAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1443–1448. http://dl.acm.org/citation.cfm?id=1642194.1642287

[29] Taavi T. Taijala, Martijn C. Willemsen, and Joseph A. Konstan. 2018. MovieExplorer: Building an Interactive Exploration Tool from Ratings and Latent Taste Spaces. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Pau, France) *(SAC '18)*. ACM, New York, NY, USA, 1383–1392. https://doi.org/10.1145/3167132.3167281

[30] Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. 2013. Escape the Bubble: Guided Exploration of Music Preferences for Serendipity and Novelty. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. ACM, New York, NY, USA, 335–338. https://doi.org/10.1145/2507157.2507223

[31] Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley. 2004. A Personalized System for Conversational Recommendations. *Journal of Artificial Intelligence Research* 21, 1 (March 2004), 393–428. http://dl.acm.org/citation.cfm?id=1622467.1622479

[32] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the unknown: visualising consumption blind-spots in recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, 1396–1399.

[33] Saúl Vargas and Pablo Castells. 2013. Exploiting the Diversity of User Preferences for Recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (Lisbon, Portugal) *(OAIR '13)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, 129–136. http://dl.acm.org/citation.cfm?id=2491748.2491776

[34] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. ACM, New York, NY, USA, 351–362. https://doi.org/10.1145/2449396.2449442

[35] Rebecca M Warner. 2012. Applied statistics: From bivariate through multivariate techniques. Sage Publications.

[36] Pontus Wärnestål. 2005. User evaluation of a conversational recommender system. In *Proceedings of the 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 32–39.

[37] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction* 28, 3 (01 Aug 2018), 237–276. https://doi.org/10.1007/s11257-018-9205-x

[38] Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding User Interactions with Podcast Recommendations Delivered via Voice. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. ACM, New York, NY, USA, 190–194. https://doi.org/10.1145/3240323.3240389

[39] Jiyong Zhang and Pearl Pu. 2006. A comparative study of compound critique generation in conversational recommender systems. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*. Springer, 234–243.

[40] Mi Zhang and Neil Hurley. 2009. Novel Item Recommendation by User Profile Partitioning. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '09)*. IEEE Computer Society, Washington, DC, USA, 508–515. https://doi.org/10.1109/WI-IAT.2009.85

[41] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2124295.2124300

[42] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. ACM, New York, NY, USA, 22–32. https://doi.org/10.1145/1060745.1060754