Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences

Yu Liang
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
y.liang1@tue.nl

Martijn C. Willemsen
Eindhoven University of Technology
5600 MB Eindhoven
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
m.c.willemsen@tue.nl

ABSTRACT

Previous studies on exploration have shown that users can be nudged to explore further away from their current preferences. However, these effects were shown in a single session study, while it often takes time to explore new tastes and develop new preferences. In this work, we present a longitudinal study on users' exploration behavior and behavior change over time after they have used a music genre exploration tool for four sessions in six weeks. We test two relevant nudges to help them explore more: the starting point (the personalization of the default initial playlist) and the visualization of users' previous position(s). Our results show that the personalization level of the default initial playlist in the first session influences the preferred personalization level users set in the second session but fades away in later sessions as users start exploring in different directions. Visualization of users' previous positions did not anchor users to stay closer to the initial defaults. Over time, users perceived the playlist to be more personalized to their tastes and helpful to explore the genre. Perceived helpfulness increased more when users explored further away from their current preferences. Apart from differences in self-reported measures, we also find some objective evidence for preference change in users' top tracks from their Spotify profile, that over the period of 6 weeks moved somewhat closer to the genre that users selected to explore with the tool.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **User models**; **Interactive systems and tools**; • **Information systems** → **Recommender systems**.

KEYWORDS

Music genre exploration, Longitudinal study, Nudge, Anchoring, Preference development

ACM Reference Format:
Yu Liang and Martijn C. Willemsen. 2022. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3523227.3546772

RecSys '22, September 18–23, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9278-5/22/09.
https://doi.org/10.1145/3523227.3546772

1 INTRODUCTION

Recommender systems typically reinforce users' current preferences and are not designed for exploration purposes. True exploration requires that users step out of their filter bubble and move away from their current preferences, while recommender systems typically recommend items based on your current (and historical) preferences. Several exploration-oriented recommender systems have been developed so far [1, 2, 7, 16, 22]to encourage users to explore new tastes. However, it is not clear yet whether these exploration tools are helpful in the long run, as they are always evaluated in a single session.

To motivate users to step out of their bubble, digital nudging [3, 6] can serve as a way to encourage users to explore away from their current preferences through (re)shaping the choice architecture. In the previous work [13], we proposed an interactive genre exploration tool for users to explore new music genres. We found that default recommendation settings influenced the users' exploration behavior: users could be nudged by default recommendation settings to explore further away from their preferences. This earlier study showed that combining personalization with nudging serves as a promising approach to support exploration. However, the study had several limitations. It only measured the effect of defaults in a single session, and users did not actually experience the exploration playlist before evaluating the system.

To further evaluate if exploration is successful in influencing users' preferences and tastes, it is also necessary to measure users' exploration and listening behavior over a longer period. Such longitudinal studies are rare, but important to understand if exploration tools and nudges have a durable effect on user preferences. Additionally, the longitudinal study allows us to look into the lasting effect of nudging, i.e., how long the effect of nudging persists. In this paper we report the results of a six-week longitudinal study with over 300 users using the genre exploration tool in four consecutive sessions, starting with a playlist, by default more personalized or more genre-representative, from a self-selected genre to explore. The central question is whether initial default recommendation playlists and visual anchoring of users' previous positions influence users' subsequent exploration interaction behavior and experience with the playlist in a longer time frame, and to what extent the exploration tool helps users to move away from their current preferences into the direction of the selected genre in terms of actual listening behavior change.

2 RELATED WORK

2.1 Music exploration

In the music domain, many exploration-related tools have been proposed to help users with music discovery and develop new tastes [1, 2, 7, 16, 22]. More and more exploration tools have been focused on personalized exploration which takes users' current preferences into consideration [1, 2, 11, 12, 16, 22]. For instance, Cai et. al. designed the conversational critique-based music exploration tool [2] and found the tool was helpful to support users to explore around their current preferences. More recently, Petridis et al., [16] developed an interactive web tool "TastePaths" to help users understand the relation between music genres by visualizing the genre relation with a graph connected by artists. They found that personalized visualization built with users' top-listened artists was helpful for users to explore something new compatible with their current tastes. For evaluation, the effectiveness of these exploration tools and user experience were mainly evaluated within a single session of exploration. Few of the previous studies did a follow-up study to evaluate users' behavior and actual preference change in a longer time frame.

2.2 Digital nudging and exploration

Recommender systems can be seen as implicit digital nudges [3, 23], which reshape users' decision structure by ordering the recommendation list in a certain way (more relevant items at the top) and guide users to what they like, based on their historical preferences. However, there is a crucial difference between nudges in the classical sense that try to help people to make (better) decisions but are not personalized and recommendations that inherently try to incorporate user knowledge into what is advised. In that sense, Starke et al. showed that nudging on top of personalized recommendations often only shows a limited effect on users' choices [19, 20]. However, in the case of exploration, nudges might be very helpful for exploration [6] and move users away from their current preferences, for example, by visualizing users' blind spots [9, 24].

In their review paper, Jesse and Jannach [6] describe a taxonomy of nudging mechanisms, classifying them as nudges that restructure/change Decision Information (such as changing information salience), Decision Structure (defaults, ranking), Decision Assistance (reminders) and Social Decision Appeal (such as using social norms). In our recent work on music genre exploration [13], we applied several of these techniques to support users in genre exploration, mostly employing Decision Structure nudges. In the genre selection interface, users were nudged to explore further away by reordering the genre selection list with more distant genres presented at the top and compared this against an order in which the closer genres were on top. Moreover, a genre was pre-selected in the list as a default option. In this earlier work, we found that users tended to select the closer genres to explore, but they were nudged to explore more distance genres with the more distant genres first in the list. However, more experienced users, scoring higher on the Musical Engagement scale of the Music Sophistication index (MSI) [15], were less sensitive to these nudges. Another nudging mechanism employed was testing different default positions of a trade-off slider. The slider allowed users to adjust the recommendations from the most representative songs of a music genre (songs representing the mainstream tastes of a certain genre) to the most personalized songs within the genre. The slider was either defaulted to the most representative, most personalized, or in the middle position. This earlier work showed that default slider positions indeed affected the final slider position, which further influenced the absolute distance of the resulting exploration playlist from users' current preferences.

However, this previous study was conducted within a single session. We did not measure the lasting effect of nudging. Understanding how long nudging lasts may uncover more about whether nudging for exploration is beneficial in the long run. Additionally, we also did not measure how users actually experienced the playlist after listening to it: we measured only their perceptions in terms of the helpfulness, but not the final satisfaction they would have with the generated playlist and their actual preference change in the long run. Given that preference development needs a long time, it is also crucial to explore the effectiveness of the exploration tool on users' experience and preference change in a longer time frame than just one session. As exploration is not an easy task, in this work, we also further explore additional nudging mechanisms [10] to make users aware of their previous slider positions by anchoring this position in the visualization (A Decision Information nudge). In this way, users might be more aware of where they come from and how the new playlist relates to the earlier ones they experienced. We expect anchoring users to their previous position would be especially useful when they are allowed to interact with the exploration tool in several consecutive sessions.

2.3 Longitudinal study

There is some work done in human-computer interaction about how to evaluate the user interface over time [4], for example on the change of users' perceived usability of the interface. In the context of recommender systems, longitudinal studies are not common, as recommender systems often consider behavior initiation rather than how behavior is consumed and maintained in the long run [17], with only a few exceptions. Taijala et al [21] measured users' experience with the movie exploration tool with a long-term measurement, whether or not they come back to the system. The work by Starke et al [18] also performed a long-term measure on users' energy-saving behavior four weeks after they received the energy-saving advice, but the authors did not measure the continuous change of users (whether users took the energy-saving advice was only measured once).

3 RESEARCH GOALS

The current study aims to extend the earlier work on music exploration with a longitudinal study that measures users' actual experience with the exploration tool and potential preference change in a longer time frame, over a period of 6 weeks in 4 consecutive sessions. With this unique longitudinal study design, we explore the effectiveness of nudges in a longer time frame than just one session. Specifically, we explore the effect of defaults in the longitudinal study to test the lasting effect of default, and visual anchors as a nudge for more effective exploration. For evaluation, we measure both subjective user experience, the intermediate user experience between sessions and final user experience, as well as objective preference change of users, potential change in preferences in terms

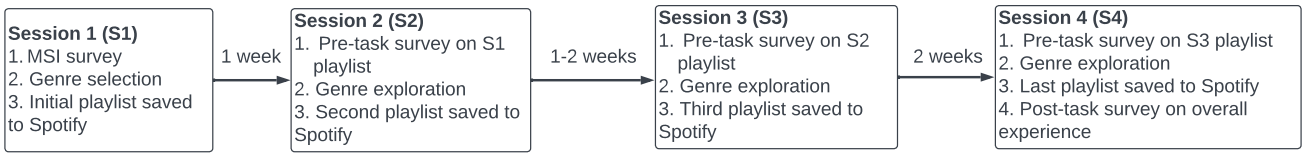


Figure 1: Longitudinal study flow

of a change in their top-listened tracks in their Spotify profile after 6 weeks. To the best of our knowledge, this study is one of the first studies conducting a longitudinal study to measure users' interactions with the recommender system (exploration-oriented) and experience with recommendations over time in multiple consecutive sessions.

The first goal of the study is to explore the pure effect of default recommendations on users' initial exploration experience. We expect that the default settings play a role in users' initial exploration experience, which subsequently shapes their future exploration. Therefore, our first research question is mainly focused on users' initial experience after Session 1. As shown in Figure 1, Session 1 asks users to select a genre to explore for the coming weeks and then provides them with either a highly personalized playlist or a highly representative playlist from that genre to listen to in the following week. After a week of interacting with the playlist, users are invited to Session 2 and fill in a pre-task survey on Session 1 playlist (about their engagement, interaction and experience with the playlist). Furthermore, we also subsequently explore how default settings and the initial listening experience influence users' interaction with the exploration system (e.g., slider usage) in Session 2. Our first research question is: *RQ1: How do default recommendations (personalized versus more representative) influence users' initial exploration experience and subsequent interaction with the exploration system?*

The next two sessions (Session 2 and Session 3) aim at exploring the genre exploration process further. To examine the lasting effects of defaults, we further explore to what extent the default recommendations from the initial session influence users' further explorations. Furthermore, we wonder if users keep exploring over time and if there is a general pattern in terms of the trade-off slider position that stabilizes. Do users slowly migrate towards more representative settings if they find the exploration useful to find more genre-specific songs, or on the opposite, do they move more towards more personalized settings if they get less interested in the genre they explore? We also look into the change in their perceived helpfulness, personalization and satisfaction in each session and explore how users' perceptions relate to their (self-reported) interactions with the playlist in between sessions. Our second research question is, *RQ2: How do user interaction with the system (slider usage), experience and exploration behavior change across subsequent sessions, and does any effect of defaults and visual anchors persist or fade away across the sessions?*

After 4 to 6 weeks, we engage participants in the final session (Session 4). To understand users' overall experience with the system, we measure their overall user experience with the music genre exploration tool (in terms of usefulness, perceived control and helpfulness) and their overall system usage. Moreover, we also look into users' actual preference change over time by comparing their current Spotify listening preferences (top-listened tracks) with their initial Spotify listening preferences (measured in Session 1). This final session allows us to provide answers to our last research questions.

RQ3a: How do users evaluate the genre exploration tool after several interactions in terms of user experience, and how does this differ with their overall system usage and nudges (the two default manipulations and visual anchoring)?

RQ3b: Does genre exploration result in actual preference change (in terms of users' top-listened tracks on Spotify), and how does this depend on users' overall system usage and nudges (the default manipulations and visual anchoring)?

4 STUDY DESIGN

4.1 Recommendation dataset and methods

For the study, we adopted the genre dataset used in previous studies for recommendations [13]. The genre dataset contains the genre typical tracks of 13 music genres[1]. The dataset was constructed from the genre-typical artists from Allmusic.com[2] and then enriched with Spotify API[3]. In this dataset, each genre is associated with a set of artists and tracks, with each artist associated with a list of tags (more fine-grained genres) and each track represented by a set of audio features. To generate personalized recommendations, we rely on Spotify API to get users' current musical preferences, i.e., their top-listened artists and tracks.

4.1.1 Genre recommendation. In Session 1, users needed to select a genre from the genre list to explore before receiving the initial playlist. In the previous study [13], it was shown that putting more distant genres at the top of the list nudged users to select a more distant genre to explore. To nudge users to explore, in this study, music genres were also presented in the order of relevance ascending, so that genres that were more distant from users' current preferences were put at the top of the list. Following the previous study, we calculated the user-genre relevance score with the Personalized PageRank algorithm [5]. With the genre dataset, we created a tag network using the tags associated with all artists: an edge was created between two tags if they co-occurred (associated with the same artist), and the weight of the edge was set to the number of their co-occurrences. The user-genre relevance was calculated as the likelihood of a random walk from the users' current preferences

[1]The music genres are: Avant-garde, Blues, Classical, Country, Electronic, Folk, Jazz, Latin, New-age, Pop/Rock, Rap, Reggae, and R&B
[2]https://www.allmusic.com/genres
[3]https://developer.spotify.com/documentation/web-api/

(represented by the tags associated with the top-listened artists) to the genre (represented by the tags associated with the artists within the genre). For details of the genre recommendation algorithms, we refer readers to the earlier work [13].

4.1.2 Track recommendation. Following the earlier work [11, 13], we also adopted the track recommendation method (a simple content-based approach) to recommend tracks within the selected genre. Here we only briefly describe the algorithm, and readers are recommended to refer back to earlier work for details and formulas. Users' preferences are modeled by their top-listened tracks with Gaussian Mixture Model in the four audio feature dimensions (retrieved from Spotify API): energy, valence, danceability and acousticness. Note that even though Spotify API provides more audio features, these four dimensions were selected for two reasons: (1) these features are sufficient to describe songs genres in terms of arousal, valence and depth [14], and (2) these features represent the important dimensions most genres differ on [12]. Given a music genre, the personalized recommendation method maps each candidate track (within the selected genre) based on users' preference model in each feature dimension, then averages the matching score over the feature dimension, and at last returns the tracks matching best with the user's current preferences. The representative recommendation method maps each candidate track with the genre-typical tracks and returns the tracks matching best with the genre-typical tastes. To balance personalization and representativeness of the recommendations, the two methods can be combined based on a personalized weight: $score_{comb} = w * score_{pers} + (1-w) * score_{repre}$. In the user study, w is the personalization level users set the recommendations with the slider (see Figure 2 (b): the trade-off slider).

4.2 Conditions

In the study, two factors were manipulated between subjects: (1) default recommendation lists and (2) visualization of previous positions. In Session 1, users were randomly assigned to a more representative (w set to 0.2) or a more personalized initial playlist (w set to 0.8). After completing the first session, users were then randomly assigned to one of the two visualization conditions, i.e., whether or not they would be able to see their previous explored position(s) in the visualization graph (the purple circle in Figure 2 (d)) in Session 2 - 4. The study consisted of four sessions allowing for a detailed comparison of individual user interactions with the system, listening behavior, user experience, and musical preference change over the four sessions.

4.3 User Interface

Figure 2 shows the exploration interface[4] of an example user. The basic elements of the interface are: (a) recommended tracks presented in a playlist style, (b) the "trade-off" slider for adjusting the recommendation personalization level, from the most representative to the most personalized, and (d) the contour plot visualization for showing the relation between the genre (green), the user's current preferences (orange) and the recommended tracks (blue triangles) in the four audio feature dimensions[5]: energy (calming to exciting), valence (negative to positive), acousticness and danceability. The contour plot visualization was adapted from the previous work [12], in which the contour plot was found to be helpful to improve users' understanding and perceived helpfulness over the recommendations during genre exploration. The visual anchor (purple circle) representing the position of users' previous playlist was only available in the visual anchoring condition. Additionally, the *show exploration history* button (element (c)) was available after Session 2 for users in the visual anchoring condition. They could switch it on (the button is off by default) to check their exploration history in the visualization.

4.4 Study Procedure

The four sessions of the longitudinal study were spaced out across six weeks between February 2022 and March 2022. Each session took around 5 to 7 minutes to complete. The study was approved by the ethical board of Eindhoven University of Technology[6].

Figure 1 shows the flow of the longitudinal study. In the first session, participants first needed to agree to the informed consent, and then logged in with their Spotify account. Next, they were asked to fill in a survey about their musical sophistication on Active Engagement and Emotional engagement [15]. After the survey, they landed on the genre selection phase, in which they needed to "select a new music genre from the genre list to explore", with the genre list sorted by the most distant genres first (the genre list was presented on a single page). In this phase, participants were also told that "for the following sessions (weeks), you will explore this selected genre, so pick a genre that you are curious about and that you would like to know better". As the final step, participants received the initial playlist (either more representative or more personalized based on conditions) and had it saved to their Spotify account. At the end of Session 1, they were given the task to listen to the playlist in the next week and were informed that the next session would be in one week.

For Session 2-4, participants who completed Session 1 were invited back to the system. After logging in with their Spotify account, they first needed to fill in a questionnaire about the previous playlist. The questionnaire also asked about their playlist listening behavior[7]. Next, they were given the task to adapt the previous playlist to what they would like to listen to in the next week, using the "trade-off" slider and the visualizations, as shown in figure 2. After adjusting the slider and generating a new playlist, participants clicked *Continue* to save the playlist to their Spotify account. Similar to the first session, participants were again asked to listen to this new playlist in the coming week (before the next session) and were informed that the next session would be in about one to two weeks. The study ended after Session 4. By the end of session four, participants were asked about their overall experience with the system (see Section 4.6 for details).

[4] In Session 1, users only received the playlist. The exploration interface is only available from Session 2.

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

[6] The study follows GDPR regulations

[7] We asked for users' self-reported behavior as this is not available from Spotify API

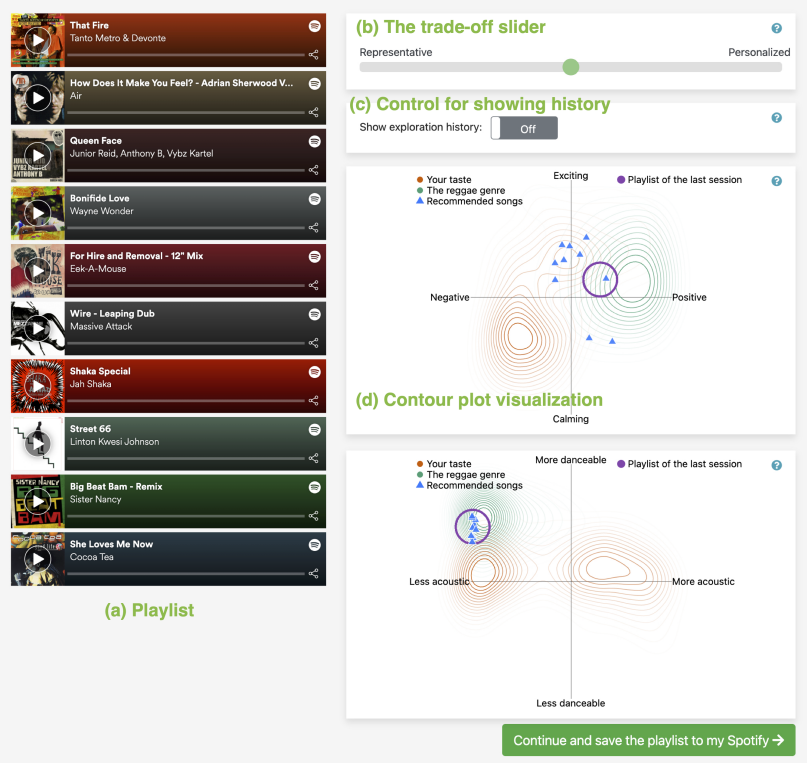


Figure 2: Genre exploration interface

4.5 Participants and payment

Participants were recruited from Prolific[8]. We first ran a short screening survey to find target participants with active Spotify accounts. During the screening survey, participants were asked four questions about (1) whether they have a Spotify account (free or premium), (2) if they share their Spotify account with others, (3) how often they listened to music on Spotify in the past year, and (4) if they are interested in a follow-up Spotify-related study that will take four sessions (with a short introduction about the longitudinal study). Additionally, participants had to be fluent in English, have a normal or corrected-to-normal vision and have no hearing difficulties. To ensure participant quality, we also limited participants to those with 95% of the approval rate and over 20 submissions. Based on the screening survey, we then invited only participants with an active Spotify account (those who listened to music on Spotify 2-5 times a week or almost every day in the past year) to the longitudinal study to ensure that sufficient listening preferences could be retrieved from their Spotify account. Participants were paid £1 per session for the longitudinal study, which was according to the payment guidelines of Prolific.

4.6 Measurements

Both subjective measurements and objective measurements were used to evaluate the change of users' exploration behavior and preference change over time. To inspect users' exploration behavior with the system over time, we recorded their interaction behavior in each session when using the system, including their tracking interaction frequencies, slider usage (available in Session 2 - Session 4), usage of the exploration history button that showed the historical changes between sessions (only available for the visualization condition in Session 3 and 4). The slider position set by users in each session was used as an indicator of exploration. However, as the slider position only shows the relative exploration distance of how much users explored from their current preferences in the selected genre, we also measured the absolute distance between users' preferences and the chosen playlist to examine the actual exploration distance, following the previous work [13]. The absolute distance was computed as the euclidean distance between the average of users' top tracks and the average of genre tracks in the audio feature space. Additionally, we also measured user-genre distance as the distance between the average of users' top tracks and tracks of the selected genre in the audio feature space.

Users' experience with the previous playlist was measured in Session 2 to Session 4 using a pre-task survey, as shown in Figure 1. Specifically, users were asked about the perceived personalization, perceived helpfulness and satisfaction of the previous playlist they listened to during the past week(s). In the pre-task survey, users also needed to answer questions about their listening behavior (*"How many times did you listen to (part of) the playlist in the last week?"* Possible options are: 0-1 times, 2-3 times and 4 times and more.), music discovery (*"Did you use the playlist to find new artists or songs of the selected genre?"*, and if so *"How did you use the playlist to*

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

find other artists or songs of the selected genre?"), and whether they favored or deleted some songs from the playlist.

Users' overall experience with the system was measured in Session 4 with a post-task survey, in which we measured users' perceived usefulness, control and perceived helpfulness of the system. Additionally, users were also asked three open questions on (*"What do you like about the music genre exploration tool?"*), (*"What do you dislike about the music genre exploration tool?"*), and (*"What other features or improvements would you suggest for the music genre exploration tool?"*). To examine the change of users' objective listening behavior on Spotify, we also recorded participants' top-listened tracks and artists on Spotify (up to 150 artists and tracks), recently played tracks (up to 50 tracks), what users saved in Spotify Library (up to 50 tracks), and artists followed (up to 50 artists) in each session.

As we are interested in whether the effects of nudging, exploration and user experience in this longitudinal setup differed for users with different musical expertise, we also measured users' musical expertise level with the Goldsmiths Musical Sophistication Index [15] on Active Engagement (MSAE) and Emotional Engagement (MSE) at the beginning of the study in Session 1. However, the analyses did not show many effects or moderations of this measure, so we do not report these in much detail further.

5 RESULTS

5.1 Demographics and genre selection

After removing those who failed the attention check, we ended up with 338 participants (265 with premium accounts and 73 with free accounts, age: 27.8 ± 8.0, gender: 187 male, 150 female, 1 unknown) who completed at least the first two sessions (Session 1 and Session 2), and 251 participants who completed all four sessions[9]. During the study, the top selected genre was *Classical*. On average, users mostly selected genres from the top of the list, i.e., more distant genres to explore (selected genre position in the list: mean=6 and median=5). Different from findings in the earlier study [13], we did not find any difference between the genre selection behavior of users with different musical expertise: users with higher musical expertise did not seem to select a closer genre than those with lower musical expertise. One potential reason could be that users, in general, made more careful decisions about what to explore in the longer term, resulting in the small difference between users with different expertise levels.

5.2 The effect of the default initial playlist

How do default recommendations (personalized versus more representative) influence users' initial exploration experience and subsequent interaction with the exploration system (RQ1)? As users selected genres that differed substantially in their distance from their preferences, we included user-genre distance as a measure in the comparison. Figure 3 shows the effect of the initial default playlist (more personalized or more representative) and user-genre distance on users' exploration behavior and user experience after Session 1. We observe that users tend to show different exploration behaviors based on their distance from the selected genre (user-genre distance), while the default initial playlists do not seem to influence users' exploration behavior much. When the selected genre was more distant (larger user-genre distance), users were more likely to listen to their initial playlist more than one time (Fig 3(a), logistic regression: $\beta = 1.008, p < .05$) and to discover new artists or songs from the selected genre (Fig 3(c), logistic regression: $\beta = 0.859, p < .05$), while they seemed to be somewhat less likely to favor songs in the playlist (Fig 3(b), logistic regression: $\beta = -0.915, p = 0.056$).

The perceived helpfulness (Fig 3(d)) and satisfaction (Fig 3(e)) depended on both user-genre distance and the type of initial playlist, while the perceived personalization did not seem to depend on either of the measures[10]. When the selected genre was close (smaller user-genre distance), users did not perceive much difference between the personalized initial playlist and the representative initial playlist. With the increase of user-genre distance, the representative initial playlist was perceived to be more helpful ($\beta = 0.614, p = 0.056$) and more satisfactory ($\beta = 1.160, p < .01$). Comparing the two playlists by means of the interaction of user-genre distance and playlist type, we see that with the increase in the user genre distance, users perceived the representative playlist to be more satisfactory and more helpful than the personalized playlist (negative interaction effects of the initial playlist's personalization level and user-genre distance: $\beta = -1.198, p < .01$ for helpfulness and $\beta = -1.206, p < .05$ for satisfaction).

Consistent with the findings in the previous work [13], the default initial playlist influenced where users put the slider in the next session. The more representative default nudged users to explore further away from their current preferences by making them put their slider at a less personalized level in Session 2 than the more personalized initial playlist (Fig 3(f), $\beta = -0.080, p < .05$). This indicates that users can be nudged by the default playlists and that this effect remains after interacting with the playlist for a week.

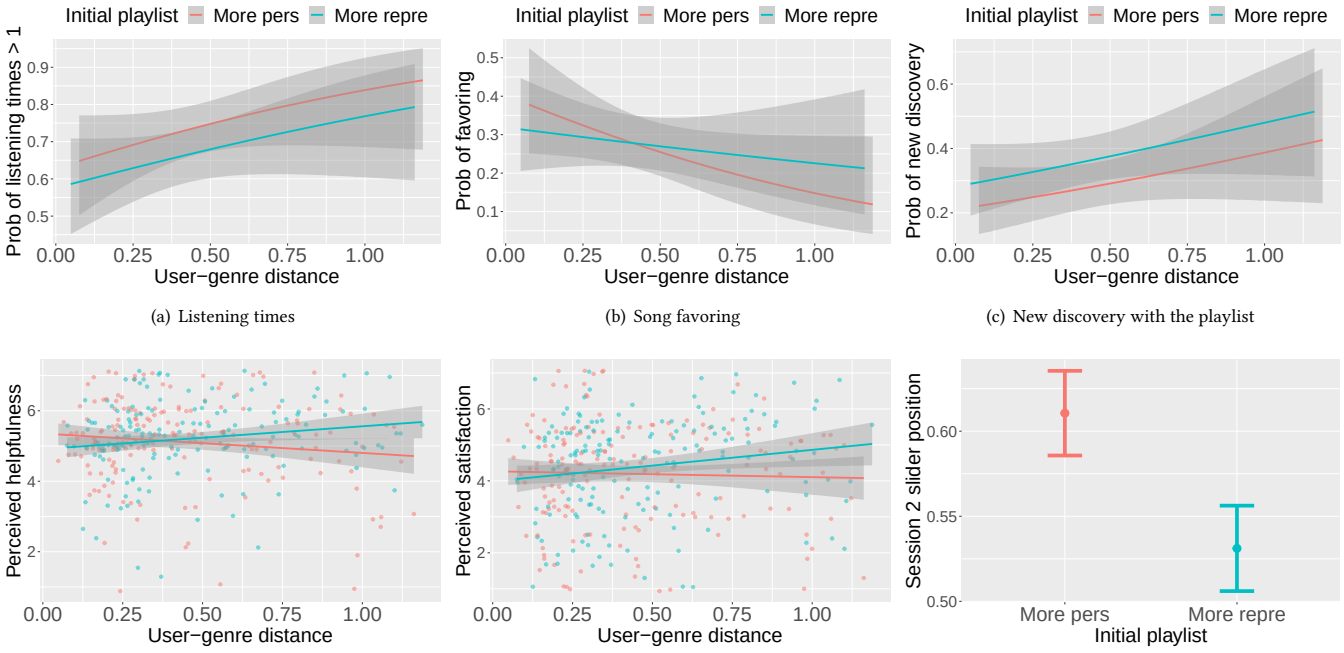
5.3 Progression over time in Sessions 2-4

5.3.1 Progression of user exploration behavior during tool usage. How does user interaction with the system (slider usage), experience and exploration behavior change across subsequent sessions, and does any effect of defaults and visual anchors persist or fade away across the sessions (RQ2)? For this research question, we look into users' exploration behavior during tool usage in the subsequent sessions. Additionally, we wonder whether users whether the default initial playlist still showed an effect on users' exploration behavior in later sessions, and whether the visualization was helpful to anchor them to their previous positions. Slider usage frequency depended on the visualization of the previous positions and decreased over sessions (Figure 4(a)). When their positions from previous sessions were visualized, users interacted with the slider somewhat more (conditions PersVis and RepreVis) ($\beta = 1.281, p < .05$)[11] than if not. As they became familiar with the tool and the recommendations, they seemed to use the slider less across sessions (negative linear effect across sessions: $\beta = -2.930, p < .001$). After Session 2, users

[9] Several less engaged users from the first two sessions who spent very limited time with the tool were not invited for the last 2 sessions

[10] Note that the results were confirmed by a SEM model for the first session, but we only report the simple results for each scale separately for brevity.

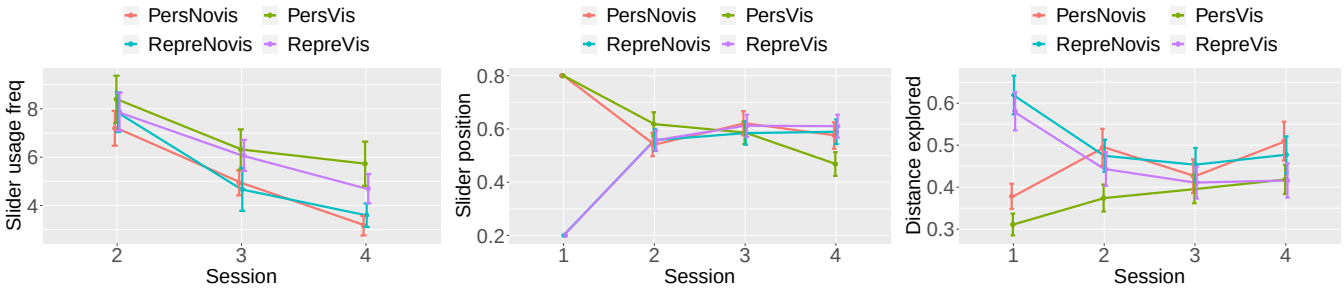
[11] The regression coefficients are from a multilevel model accounting for the repeated measurements across the sessions



(a) Listening times (b) Song favoring (c) New discovery with the playlist

(d) Perceived helpfulness (e) Perceived satisfaction (f) Next slider position

Figure 3: Reported listening behavior and experience after Session 1. In (d) and (e), the scatter plot shows the original distribution (jittered), with the shaded area indicating 95% confidence interval.



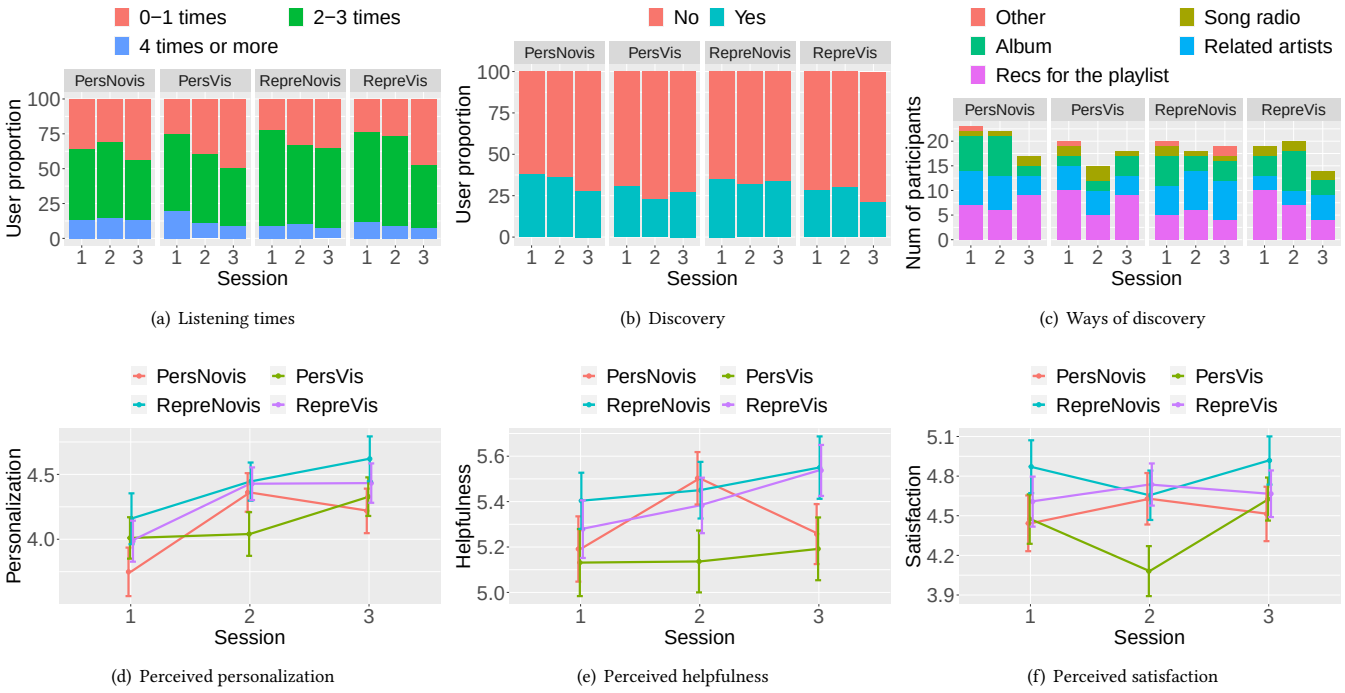
(a) Slider usage frequency (b) Slider position (c) Distance explored

Figure 4: User exploration behavior during tool usage over sessions across the two default conditions (Repre is the representative default and Pers is the personalized default) and visualization of the history (Vis versus Novis) Error bars indicates standard errors.

seemed to put the "trade-off" slider[12] somewhere in the middle and explored around. As shown in Figure 4(b), the default initial playlist does not seem to influence strongly where users put their slider in the second or later sessions or how much they explored away from their current preferences (measured as the absolute distance between users' top tracks and tracks in the generated playlist). Surprisingly, we also do not observe any anchoring effects of visualization on users' exploration behavior: visualization of their previous positions did not make users stay closer to their last slider position.

5.3.2 Progression of reported listening behavior. Figure 5 shows the reported listening behavior of users and their experience with the playlist received in each session. Note that the reported listening behavior and user experience were not evaluated immediately after each session but at the beginning of the next session, so users were given around one week to listen to the playlist (as explained in Section 4.4 and Figure 1).

[12]Note that the slider was available from Session 2.



(a) Listening times (b) Discovery (c) Ways of discovery

(d) Perceived personalization (e) Perceived helpfulness (f) Perceived satisfaction

Figure 5: Reported listening behavior and user experience over sessions. In (d) to (f), the error bar indicates the standard error.

As shown in Figure 5(a), more than 50% of the participants reported that they listened to the playlist at least two times. The proportion is the highest for Session 1 (86%) and lowest for Session 3 (55%), showing that on average users listened to the generated playlist less frequently over time. This effect is also confirmed by a multi-level logistic regression which predicts the probability of listening more than once based on sessions. Comparing to Session 1, users were less likely to listen more than once in both Session 2 ($\beta = -0.469, p = 0.073$) and Session 3 ($\beta = -1.376, p < .001$). Figure 5(b) shows that more than 25% of the participants reported that they used the playlist to discover something new, and again this proportion is the highest for Session 1 (32%) and lowest for Session 3 (27%). This effect is not significant in the multi-level logistic regression which predicts the probability of discovery (Fig 5(b)). Compared to Session 1, users seemed to be less likely to discover something new in Session 3 although the difference is not significant ($\beta = -0.426, p = 0.086$). As shown in Figure 5(c), the top three ways for users to discover new artists or songs of the genre are (1) through Spotify recommendations based on what's in the playlist, (2) through related artists, and (3) through songs of the same album. Around 25% of the participants reported that they favored at least one song in the playlist, and the proportion does not differ across sessions. On average, around 5% of the participants reported that they deleted at least one song in the playlist, and surprisingly this proportion is the lowest for Session 1 (1.1%) and highest for Session 3 (9.5%). This might indicate that some participants became more critical over time and started to adjust the playlist if there were songs they did not like much. Note that we did not find much difference in listening behavior across the defaults and visualizations. None of these factors were significant predictors in the multilevel regression models.

5.3.3 Progression of user experience. Figure 5(d) to Figure 5(f) shows the perceived personalization, helpfulness and satisfaction of participants on each session's playlist. Over time, users perceived the playlist to be more personalized and more helpful, while the perceived satisfaction did not seem to differ across sessions. We tested these differences using multilevel regressions across the 3 sessions. Comparing to Session 1, both the Session 2 and Session 3 playlists were perceived to be more personalized ($\beta = 0.340, p < .001$ for Session 2 and $\beta = 0.422, p < .001$ for Session 3) and more helpful ($\beta = 0.116, p = 0.087$ for Session 2 and $\beta = 0.134, p < .05$ and Session 3). On average, users' perceived personalization, helpfulness and satisfaction did not differ much across conditions except for Session 2: users with the more personalized initial playlist and visualization of their previous position seemed to perceive the playlist from Session 2 to be less personalized ($\beta = -0.738, p < .05$) and satisfied ($\beta = -0.925, p < .05$) than users in the other conditions, as shown by the green line that is somewhat lower in Figure 5(d) and Figure 5(f). This might be related to the somewhat more personalized slider positions and lower absolute distance explored in this condition as shown in Figure 4, though these effects were not significantly different.

5.3.4 Summary conclusion of progression in Session 2-4. In summary, users' listening behavior and their experience with the playlist paint a diverse picture of results. On the one hand, default effects fade relatively quickly and users seemed to adjust the slider to a

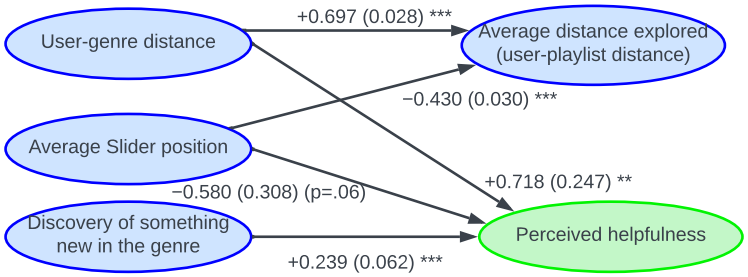


Figure 6: Structural Equation Model on users' overall experience with the system. Significant levels: ***p < .001; **p < .01; *p < .05. Numbers on the arrows indicate coefficients and standard error (in brackets) of the effect.

preferred level given the genre exploration tool, and therefore perceived the generated playlist to be more personalized and helpful over sessions. The visualization condition also seems to have a limited effect on slider position and thus on the distance explored. However, these graphs show averages across all users, we did find that slider positions within a user varied quite strongly between sessions and that users tried out different positions before ending around in the middle (at around 0.6), as shown in Figure 4. In terms of interaction behavior, we do see a decline over time in slider usage, listening to the generated playlist and using the playlist for discovery. However, even in Session 3, most users still seemed to be engaged, with more than half listening to the playlist more than once. However, we have to realize that this is self-reported (not actual usage) data. Together this makes us wonder how users evaluate the system in the end and if the exploration had any effects on their Spotify profiles after the period of 6 weeks using the tool, which we will analyze in the next section.

5.4 Overall Experience with the System

RQ3 asks about how users evaluated the overall experience and the actual impact of exploration on preference development: if users' Spotify profiles were affected by exploration. Following the user-centric evaluation framework by Knijnenburg et al. [8], we modeled the perceived helpfulness, control and usefulness in a structural equation model. Since those variables were highly correlated and most effects loaded on helpfulness, we focus the discussion on the helpfulness of the system for supporting music genre exploration (Figure 6) and how perceived helpfulness was influenced by several metrics of user behavior. Since slider position varied a lot within users, we took the average slider position across sessions 2 and 3 as a measure of reflecting how much each individual user aimed to explore across the entire period. We related these to the initial user genre distance and aggregated listening behavior across the sessions. From the listening behavior, only the factor *discovering something new* was significant.

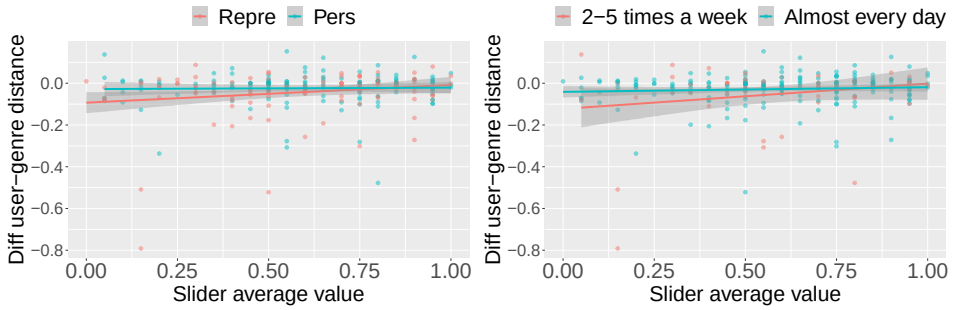
Consistent with the previous work [13], the average distance users explored (across Session 2 and Session 3) when using the exploration tool (user-playlist distance) was positively influenced by the user-genre distance and negatively influenced by the average slider position (personalization level set by the slider). Users' overall perceived helpfulness of the system was not influenced by either the default initial playlist or showing the visualization of their previous position(s). Users perceived the system to be more helpful when they explored a genre further away and put the slider at a less personalized position. Additionally, the perceived helpfulness of the system was also positively related to their music discovery in the selected genre: the more users utilized the generated playlists to discover new songs or artists in the selected genre, the more they perceived the system to be helpful.

5.5 Preference change in Spotify listening data

Apart from the self-reported measures and subjective experience, we also look into the objective change of users' top-listened tracks after several weeks of exploration, to see whether the genre exploration tool is helpful for developing new tastes[13]. This is done by comparing the user-genre distance (i.e., the distance between users' top tracks and tracks of the selected genre) measured at the beginning of Session 1 and at the beginning of Session 4 (six weeks after).

Our results show that the user-genre distance is smaller in Session 4 than in Session 1 (paired one-tailed t-test, $t = -5.4433, p < .001$, diff= -0.0339), indicating that the genre exploration tool is helpful to drive user preferences somewhat into the direction of the selected genre over time. We also find the difference was affected by the default conditions and the average slider value users set during the sessions. The default initial playlist influences somewhat how much users explored the selected genre: users explored towards the selected genre more when their default initial playlist was more representative (the blue line is always above the red line in Figure 7(a), $p < .05$). Users seemed to explore towards the selected genre less when they put the slider (on average) at a more personalized level (as also shown in Figure 7(a), marginal significant $p = 0.07$) and this effect is mostly there for those users that started at the representative default and less so for those that started at the personalized default.

[13]Note that although the daily listening data of users is a better measure for tracking their preference change, the data is not accessible through Spotify API. The retrievable listening data is restricted to users' top-listened artists and tracks.



(a) The effect of the initial playlist

(b) The effect of Spotify usage

Figure 7: Users' actual preference change. The scatter plot shows the original distribution, with the shaded area indicated 95% CI

We also checked if users' overall Spotify usage behavior affects the change in their Spotify listening data. Since we try to show preference changes towards the selected genre in users' Spotify profiles, we would expect that more frequent users might listen more to other music besides the genre exploration playlists than less frequent users. Indeed we find that the more frequent users showed smaller changes in their distance towards the genre (Figure 7(b), positive effect of more frequent users on *diff user-genre distance*: $\beta = 0.08, p < .05$) and the difference was not affected by the average slider position. The less frequent users showed larger differences, especially when their average slider position was more towards a genre-representative setting. In other words, if users explored closer to the genre, they showed larger differences in changes towards the genre (marginal significant negative interaction of user type on the difference makes the difference larger: $\beta = -0.09, p = 0.09$).

6 CONCLUSION AND DISCUSSION

Our longitudinal study provides important insights into how exploration progresses over time and how digital nudges like defaults and visual anchors can support the process. Answering the first research question, we find that a more representative (relative to a more personalized) default did increase perceived helpfulness and satisfaction after users listened to the playlist for a week, especially when they had selected more distant genres to explore. Listening behavior was not much different between default conditions, but the defaults did influence the slider position users set in Session 2. In subsequent sessions (answering RQ2), the effect of the defaults faded away, and the effects of visual anchors were also limited. Users explored a lot but on average seemed to end up with similar slider positions and exploration distance during the subsequent sessions, while perceived more personalization and more helpfulness over time. Initial engagement with the playlist (in terms of self-reported listening behavior) did drop somewhat in later sessions, but overall users also seemed to interact with the playlist substantially in later sessions.

In the final session, we measured users' overall experience with the system (RQ3a) and we checked their music preference change with their Spotify listening preferences after 6 weeks (RQ3b). We find that users perceived the system to be more helpful when the average slider position in Session 2 and Session 3 was set to a more representative level. Additionally, helpfulness increased when users had chosen a more distant genre and reported that they used the playlist to discover new items. Analyzing users' Spotify profile data after 6 weeks of interacting with the exploration tool, we find on average, user profiles did move somewhat towards the chosen genre. We also find a residual effect of the initial defaults (initial recommendation list): the Spotify profile change was larger for users with the representative initial playlist, especially when those users also set slider positions at an average more representative level. The results also show that very frequent (daily) users of Spotify showed less change in their Spotify profile toward the new genre. This is not surprising since these users will listen to much more other music than just the exploration playlist. One of the limitations of the current measure on users' Spotify profile change is that the change was measured at a coarse level (users' top-listened tracks), and it is not possible to tell whether the change was brought by the tool or by any other external factors.

In conclusion, exploring the effects of the music genre exploration system in a longitudinal study across 6 weeks and 4 sessions, we see that a personalized trade-off slider for exploration allows users to explore a genre effectively, especially for those that explore more towards the representative slider positions. The effect of nudge (default initial playlist and visual anchors) does not seem to show a lasting effect across sessions in terms of users' interaction with the exploration system (slider usage) and user experience. However, the residual effect on the change in users' Spotify profiles does seem to show that the representative default still might have triggered some users to explore a bit further away from their existing preferences, showing potential benefits of combining nudging with personalization in exploration tools even for helping users to explore in a long run.

REFERENCES

[1] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights. In *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*. ACM Press, New York, New York, USA, 35. https://doi.org/10.1145/2365952.2365964

[2] Wanling Cai, Yucheng Jin, and Li Chen. 2021. *Critiquing for Music Exploration in Conversational Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 480–490. https://doi.org/10.1145/3397481.3450657

[3] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–15. https://doi.org/10.1145/3290605.3300733

[4] Marc J. M. H. Delsing, Tom F. M. ter Bogt, Rutger C. M. E. Engels, and Wim H. J. Meeus. 2008. Adolescents' music preferences and personality characteristics. *European Journal of Personality* 22, 2 (mar 2008), 109–130. https://doi.org/10.1002/per

[5] T.H. Haveliwala. 2003. Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering* 15, 4 (2003), 784–796. https://doi.org/10.1109/TKDE.2003.1208999

[6] Mathias Jesse and Dietmar Jannach. 2021. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* 3 (2021), 100052. https://doi.org/10.1016/j.chbr.2020.100052

[7] 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* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 19–30. https://doi.org/10.1145/2856767.2856780

[8] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (oct 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[9] Jayachithra Kumar and Nava Tintarev. 2018. Using visualizations to encourage blind-spot exploration. *CEUR Workshop Proceedings* 2225 (2018), 53–60.

[10] Elisabeth Lex, Dominik Kowald, Paul Seitlinger, Thi Ngoc Trang Tran, Alexander Felfernig, and Markus Schedl. 2021. Psychology-informed recommender systems. *Foundations and Trends in Information Retrieval* 15, 2 (15 July 2021), 134–242. https://doi.org/10.1561/1500000090

[11] Yu Liang. 2019. Recommender System for Developing New Preferences and Goals. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 611–615. https://doi.org/10.1145/3298689.3347054

[12] Yu Liang and Martijn C. Willemsen. 2021. Interactive Music Genre Exploration with Visualization and Mood Control. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 175–185. https://doi.org/10.1145/3397481.3450700

[13] Yu Liang and Martijn C. Willemsen. 2021. *The Role of Preference Consistency, Defaults and Musical Expertise in Users' Exploration Behavior in a Genre Exploration Recommender*. Association for Computing Machinery, New York, NY, USA, 230–240. https://doi.org/10.1145/3460231.3474253

[14] Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. 2018. Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 101–109. https://doi.org/10.1145/3209219.3209223

[15] 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 (feb 2014), e89642. https://doi.org/10.1371/journal.pone.0089642

[16] Savvas Petridis, Nediyana Daskalova, Sarah Mennicken, Samuel F Way, Paul Lamere, and Jennifer Thom. 2022. TastePaths: Enabling Deeper Exploration and Understanding of Personal Preferences in Recommender Systems. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 120–133. https://doi.org/10.1145/3490099.3511156

[17] Alain Starke. 2019. *Supporting energy-efficient choices using Rasch-based recommender interfaces*. Ph.D. Dissertation. Industrial Engineering and Innovation Sciences. Proefschrift.

[18] Alain Starke, Martijn Willemsen, and Chris Snijders. 2017. Effective User Interface Designs to Increase Energy-Efficient Behavior in a Rasch-Based Energy Recommender System. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 65–73. https://doi.org/10.1145/3109859.3109902

[19] Alain Starke, Martijn Willemsen, and Chris Snijders. 2021. Promoting Energy-Efficient Behavior by Depicting Social Norms in a Recommender Interface. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 30 (Aug. 2021), 32 pages. https://doi.org/10.1145/3460005

[20] Alain D. Starke, Elias Kløverød Kløverød Brynestad, Sveinung Hauge, and Louise Sandal Løkeland. 2021. *Nudging Healthy Choices in Food Search Through List Re-Ranking*. Association for Computing Machinery, New York, NY, USA, 293–298. https://doi.org/10.1145/3450614.3464621

[21] 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)*. Association for Computing Machinery, New York, NY, USA, 1383–1392. https://doi.org/10.1145/3167132.3167281

[22] 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)*. Association for Computing Machinery, New York, NY, USA, 335–338. https://doi.org/10.1145/2507157.2507223

[23] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

[24] 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* (Pau, France) *(SAC '18)*. Association for Computing Machinery, New York, NY, USA, 1396–1399. https://doi.org/10.1145/3167132.3167419