Personalized Recommendations for Music Genre Exploration

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

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

ABSTRACT

Most recommender systems generate recommendations to match the user's current preference. However, users sometimes might have the goal to develop new preferences away from their current preference and use the recommender to guide them towards it. In this paper, we asked users to select a new genre to explore and studied what kind of recommendations would be more helpful for users to start exploring this new music taste. Three different recommendation methods are tested: one non-personalized which recommends the most representative tracks of the genre, one personalized method which considers songs from the new genre that best matches users' current preferences, and one mixed method which makes a trade-off between the two approaches. A comparative design was used in a user experiment in which participants were asked to evaluate the differences between the personalized method/mixed method and the non-personalized baseline. The mixed method results in recommendations that are more accurate and representative for the new genre than the personalized method. Users' perceived helpfulness for exploring the new genre is positively related to both perceived accuracy and perceived representativeness of the recommended items. Besides, recommendations from the mixed method are perceived more helpful for users high on Musical Sophistication Index for Active Engagement (MSAE). To our knowledge, this is one of the first studies using a recommender system to support users' preference development, and provides insights in how recommender systems can help users attain new goals and tastes.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Personalization**; **Recommender systems**; • **Human-centered computing → User studies**.

KEYWORDS

content-based music recommendation; user-centric evaluation; personalization; preference developing; exploration; user goals

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3320455

ACM Reference Format:
Yu Liang and Martijn C. Willemsen. 2019. Personalized Recommendations for Music Genre Exploration. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3320435.3320455

1 INTRODUCTION

Most recommender systems model users' long term preferences and provide recommendations that are close to the user's current long-term preference, such as the widely used collaborative filtering recommender system [8]. However, this type of recommendation has some limitations. For example, the user's short-term situational needs can be different from her long-term preference [9, 32]. Taijala et al. [32] developed an interactive exploration tool to help users to pick up a movie that is different from what they typically like. Furthermore, people sometimes want to change their long-term preference or develop a new preference. Studies have shown that there is a natural drive in humans seeking for novelty and change [21]. It is also indicated that there is a spontaneous devaluation in user preference on music listening [13, 14] that users would sometimes be bored of their current preference and seek for the changes. In this case, they also need a recommender system which can help them develop new preferences with the recommendations.

How can a recommender system help people that have a goal to learn a new taste? In this paper, we will put our focus on the music field since it is common that people would like to develop new music tastes, e.g., from popular to classical, and do not know where to start. We could ask them to select a new genre and start with the most representative and popular songs of that genre, but this might not be the most effective. If the genre is far away from their current preference, that new goal might be too challenging, just like a person who wants to run a marathon should start with something that is closer to his current performance level: if a person never ran before, start with 5k, if she is already an active runner that can do 10k, we might advise her how to train for a half marathon first. This is consistent with goal theories that propose small steps should be taken to achieve a goal. Similarly, we aim to personalize the new genre selection based on the user current preferences.

Such preference-based recommendations might be perceived to be more personalized and more accurate, however, the recommendations can also be perceived to be less representative of the new genre and thus less helpful to explore the new genre. Our main research question is whether the preference-based recommendations (even though perceived less representative by the user) can better support the user to explore the new genre than the most representative and popular songs of that genre.

2 RELATED WORK

2.1 Behavior change

Behavior change always takes steps, so does developing new preferences. According to the transtheoretical model [27], there are five stages in the whole process of behavior change: precontemplation, contemplation, preparation, action and maintenance. The change from the precontemplation stage to the contemplation stage requires awareness raising. However, with the goal of new preference development, the user has already thought about the change and reached the contemplation stage. In this work, we are concerned about how to help users to explore new music genres and further develop new tastes given that they want to develop new preferences, then how to move them from the contemplation stage to the preparation stage and further the action stage.

2.2 User exploration in RS

Recently, some visualization tools have been proposed in RS to encourage users' exploration [34] by raising their awareness of preference change. These visualization tools serve a purpose to help people understand their current bubble [26] by visualizing their filter bubbles/blind spots [23–25]. According to the transtheoretical model [27], this kind of awareness rising serves the first step of their behavior change, in which they change from the precontemplation stage to the contemplation stage. However, this work didn't discuss what they can do in the next step after realizing the importance of change for developing new preferences.

Interactive exploration tools have also been designed in RS to help with user exploration. MovieExplorer [32] is an interactive movie exploration tool designed to fulfill users' short-term movie tastes by allowing them to navigate in the latent space. Users can navigate in the latent space [10, 18] in rounds until they are satisfied. The navigation is done by the positive or negative feedback the user chooses to provide to the displayed items in the current recommendation round. The recommendations in the next round will move towards the movies with positive feedback and away from the movies with negative feedback. Schnabel et al. [30, 31] implement an interface component (shortlist) to support movie exploration by reducing users' cognitive overload during decision making. It is indicated that users are more likely to explore with shortlist even after finding good items. In music recommendation, a recommender system called *MoodPlay* [1] is designed to help users to explore music in a latent mood space. With the visualization of the latent mood in a 2D mood space, users can move from one mood to another in an interactive-based way.

Rather than supporting the exploration during the recommendation process, some interactive recommender systems are designed to assist the exploration of the recommendation results by allowing users to manipulate different aspects and settings of the recommendation algorithms [5, 17] with more user control [11, 12]. LinkedVis [5] allows users to explore job recommendation results with different skill settings and check whether a new skill can bring them closer to their desired jobs. MyMovieMixer [17] combines recommendation algorithms with interactive faceted filtering interface. In MyMovieMixer, users are allowed to explore the movie recommendation results under different facet settings, which increases the transparency and user control during the recommendation.

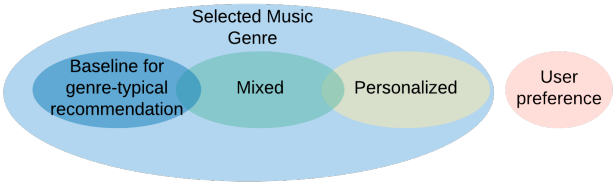

Figure 1: Different methods for new music genre exploration

A more relevant work is done by Taramigkou et al. [33], in which the shortest path from the user's current preference in the latent genre space to the target latent genre (represented by an actual target user who matches best with the latent genre) is identified in a user preference graph, with nodes corresponding to users and edges representing user to user similarity, to help with user exploration along the path. The top three listened artists matching best with the target latent genre of all users along the path will be assembled as a sequence of recommended artists for exploration. This approach suggests a sequence of artists for the user to listen to before she could reach the final latent genre by utilizing the preference of other users. However, one shortcoming is it might require the user to explore other irrelevant genres before reaching the final target genre especially when the user is far away from the target user in the user graph. Furthermore, top matched artists from the users along the path do not necessarily mean they are the good ones for preference change, as the path is only the shortest path from the user's current preference to an actual user matching best with the latent genre.

All the exploration methods discussed above start from the user's current preference. However, with the goal to explore new music genres and further develop new tastes, users might need the exploration to start directly from the new genre. Besides, they would also expect that the recommender system could help them develop the new preferences gradually in steps, guiding them towards their goals given their current preference.

2.3 Cross-domain recommendation

Recommending tracks from genres away from the user's current preference fits in the scope of cross-domain recommendation. In a cross-domain recommendation situation, recommender system has to exploit knowledge from one or more source domains to some target domain in order to perform recommendations in the target domain [6].

Similarly, we can utilize users' current preference to generate recommendations in new music genres, under the assumption that their tastes are to some extent consistent across different genres. Here, domains refer to genres. In our work, the cross-genre music recommendation will be done in a content-based way [28] with the genre-independent audio feature data from Spotify, which allows us to link users from their own genres to new genres. These audio features are sufficiently general to bridge the gap between genres. For example, we assume a user who used to enjoy classical music (usually with low energy) would also enjoy tracks with low energy

from a different genre such as rap. In this way, recommendations from the new genre can take the user's current music preference into consideration. Six high-level semantic audio features will be used for the cross-genre recommendation: acousticness (whether the given track is acoustic), danceability (how suitable the given track is for dancing), energy (the intensity of the given track), liveness (whether the given track is live), valence (the positiveness of the given track) and speechiness (whether the given track is speech-like).

2.4 Summary and hypotheses

With the goal to develop new music preferences, users would need the recommender system to help with their exploration in the new genre. The exploration will allow users to prepare for the real change of their music taste and take actions towards it (corresponding to the preparation stage and the action stage of the transtheoretical model respectively). Furthermore, a more guided exploration, taking users' current preferences into account, might better help them to explore the new music genres. Based on the above ideas, our research question is to explore whether a preference-based genre-exploration approach can better help users to explore new music genres than the non-personalized one, even though the recommendations might be perceived less representative (RQ).

To test our research question, three different methods for new music genre exploration are proposed as shown in Figure 1. The baseline genre-typical method will generate recommendations in a non-personalized way with the most typical items from the selected music genre. Two preference-based methods will give recommendations from the selected genre based on the user's current preference. The personalized method will give items from the new genre that are most similar to the current taste. The mixed method serves as a trade-off to balance the personalized and genre-typical recommendations (baseline) since a playlist that is too personalized might not be representative of the genre and thus might not be helpful in achieving the goal.

We will investigate our research question under the user-centric framework from Knijnenburg et al. [15, 16] with a comparative design as used by Ekstrand et al. [7], putting two lists generated by two methods side by side, to measure the perceived differences between methods. To answer our research question, we need to understand how perceived personalization and representativeness would affect perceived helpfulness. Besides, perceived accuracy might also play a role in perceived helpfulness, although there could be an overlap between perceived personalization and accuracy. We will also include diversity as one dependent measure as it has shown to influence perceived accuracy in earlier studies [35] and because we expect the baseline genre-typical list to be less diverse than the list from the mixed method. Finally, we expect some of these effects to be moderated by the expertise of the user, as measured by the Musical Sophistication Index [22]. User higher on musical sophistication might be better at judging the representativeness of a list for the genre, or how helpful a list is to explore a new genre. Moreover, they might respond differently to the two preference-based methods, for example, being able to cope better with the mixed method as they can easier judge how these items relate to both their own preferences as well as the genre-specific list. Our hypotheses are:

- **H1**: The perceived helpfulness will increase with both perceived representativeness of the genre as well as the perceived accuracy of the playlist
- **H2**: The recommendations from the baseline method will be perceived more representative than those from the preference-based methods, with the difference between personalized method and the baseline larger than the difference between the baseline and the mixed method
- **H3**: The recommendations from both personalized and the mixed method will be perceived more accurate and/or personalized than those from the baseline
- **H4**: The recommendations from the mixed method will be perceived more diverse than those from the baseline
- **H5**: Level of musical sophistication might moderate the effects of preference-based methods on representativeness and helpfulness.

3 METHODS

We choose Spotify as our experiment platform. Spotify has become one of the largest music streaming platforms with millions of active users in the Netherlands. Making use of the Spotify Web API [1], we are able to access a user's top tracks and retrieve a track's audio features, with which we build our content-based approach for cross-genre music recommendation (as explained in Section 2.3). To help with the new music genre exploration, we need prototypical songs from several music genres. However, to the best of our knowledge, there is no such ground truth. We therefore constructed a genre dataset ourselves.

Following the work of Schedl et al. [29], we start with a dataset of prototypical artists from a genre before constructing the dataset of prototypical tracks. Firstly, highlighted artists from several music genres are retrieved from Allmusic.com[2]. Here, we rely on the expert opinions of Allmusic.com editors on the genre prototypical artists. To enlarge the genre dataset, related artists of these highlighted artists are also retrieved with the Spotify API. So far, the highlighted artists and their related artists consist a dataset of genre prototypical artists.

However, some second-level artists are not necessarily from the genre, so we drop those unrelated artists. As the first step, we retrieve all tags associated with all artists from the dataset with the Spotify Web API. Then we build a dictionary with the tags from the first-level artists. Within each genre, tf-idf scores are computed for all tags in the dictionary to upweight the tags that occur often within the genre and downweight the tags that occur in many different genres. Top 20 tags with the highest tf-idf score are selected as the tags of the genre. Artists without the top 20 tags within the genre are dropped as unrelated artists. Finally, top tracks are retrieved for all the first-level artists and second-level related artists from Spotify. These tracks together comprise a dataset of prototypical tracks. Table 1 presents the summary of the dataset.

[1] https://developer.spotify.com/documentation/web-api
[2] https://www.allmusic.com/genres

Table 1: genre dataset

genre	# tracks	# artists
avant-garde	3307	349
blues	2489	252
classical	4116	444
country	2728	281
electronic	3818	388
folk	3101	317
jazz	3346	347
new-age	3897	395
rap	3351	345
r&b	3299	334

3.1 Musical preference modeling

According to the Spotify Web API reference, a user's top tracks are retrieved based on calculated *affinity*, a metric to measure the user's expected preference towards a particular track based on the user's behavior. Similar to the work by Bogdanov et al. [4], where a user's musical preference is modeled by a set of tracks the user thought could best show her musical preference, we model users' music preferences with their top tracks. We assume the top tracks can represent a user's musical preference well. However, rather than representing the user's musical preference with an overall model in multi-dimensional space as in the previous work [4], we model the user's musical preference towards each individual audio feature separately. The shortcoming of the the overall preference model is that it will give more weights to the matched features and the derived recommendations will keep reinforcing the user's old tastes even in new and far away music genres, while our goal is to help users develop new tastes by transferring some of her old tastes.

The user's preference towards each individual audio feature (e.g., acousticness) is modeled by the corresponding feature values extracted from her top tracks. For each feature model, a Gaussian Mixture Model is trained to represent the user's preference as a probability density function. Each Gaussian Mixture Model is initialized by k-means and trained with the EM algorithm [3] as in the previous work [4]. Information-theoretic criteria (BIC) is responsible for selecting appropriate *component numbers* (range from 0 to 10) and *covariance type* ({'spherical', 'tied', 'diag', 'full'}). The total number of top tracks we could retrieve from Spotify is at most 150 for each user (50 tracks for each time range: short-term, medium-term and long-term). Due to the small number of top tracks available, a Gaussian filtering function is added to smooth the user model.

3.1.1 Personalized method. The personalized method generates recommendations based on the user model. With the user preference model, we can map the corresponding track feature value to a probability density value for all candidate tracks. If a track feature is in a high-density area, it is more likely to be a match. To serve the recommendations, we first map the probability density value for each track on each audio feature with the corresponding preference model. Given an audio feature f, a ranked candidate track list will be returned based on their mapped density values. The ranking score of track i on feature f is computed as $score_f(i) = n - R_f(i) + 1$, in which n is the total number of candidate tracks and $R_f(i)$ represents the ranking of track i in the ranked list. The final preference-based recommendation score of track i is computed by aggregating the ranking score of all feature dimensions, which is equal to the sum of $score_f(i)$ for all the audio features: $score_{personal}(i) = \sum_{f \in F} score_f(i)$. Top 10 tracks with the highest recommendation scores will be returned as recommendations (see Algorithm 1).

The idea of the ranking score aggregation is derived from Borda count. Borda count has been widely used in group recommendation [19, 20] to aggregate recommendation results from users in the group. It computes scores for each item in a group member's ranked list based on the item's position in the list [2]. Here, Borda count is used to aggregate recommendation results from different audio features.

Algorithm 1 Personalized recommendation

$\boldsymbol{value_f(i)}$: feature value of track i on feature f; $\boldsymbol{GMM_f}$: user GMM preference model on feature f; $\boldsymbol{density_f(i)}$: density value of track i on feature f; $\boldsymbol{score_f(i)}$: ranking score of track i on feature f; $\boldsymbol{score_{personal}(i)}$: personalized recommendation score of track i; $\boldsymbol{n}$: number of tracks, $\boldsymbol{R_f(i)}$: position of track i in the ranked list of all candidate tracks, ranked by $density_f(i)$ (from high to low); $\boldsymbol{F}$: set of audio features

```
for each audio feature f do
    for each track i do
        density_f(i) ⇐ GMM_f(value_f(i)) //density value
    end for
    score_f(i) ⇐ n − R_f(i) + 1 //ranking score
end for
for each track i do
    score_personal(i) ⇐ Σ_{f∈F} score_f(i) //recommendation score
end for
return top 10 tracks with the highest score_personal value
```

3.1.2 Baseline method for genre-typical recommendation. To construct a genre-specific baseline, we build a recommendation algorithm based on how well a candidate track matches the mainstream typical taste of the genre. The typical taste of a genre is modeled by a multi-dimensional Gaussian Mixture Model [4] with the six different audio features associated with the top tracks from the first level artists. Each candidate track can be represented as a vector of the six audio features. Based on the mainstream taste of the genre modeled by GMM, a density value is returned for each candidate track. The higher the density value, the better the track matches the mainstream taste. Finally, the top 10 tracks with the highest density values will be returned as recommendations.

3.1.3 The mixed method. The mixed method combines the recommendations from the personalized method and genre-typical baseline, which serves a trade-off to balance the personalized taste and mainstream taste of the genre. In the mixed method, a ranking score is firstly computed for each candidate track based on its position in the final ranked list generated in the two methods ($rscore_{personal}(i)$ and $rscore_{base}(i)$). The mixed score is then

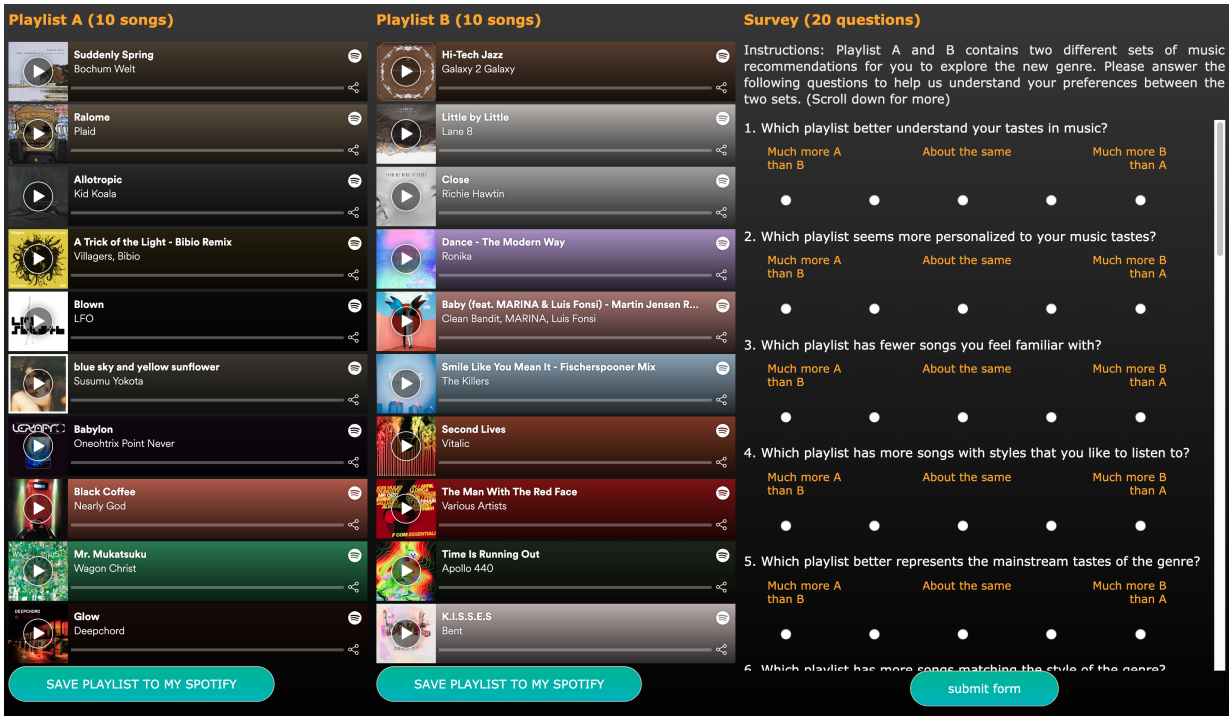

Figure 2: Screen shot of the experiment interface

computed as the weighted sum of the two scores: $score_{mix}(i) = w * rscore_{personal}(i) + (1 - w) * rscore_{base}(i)$, in which w is the weight set to the personalized method (in the experiment, $w = 0.5$). Top 10 tracks with the highest mixed scores will be returned as recommendations, as shown in Algorithm 2.

Algorithm 2 The mixed method

$\boldsymbol{r_{personal}(i)}$: position of track i in the final ranked list generated in personalized method; $\boldsymbol{r_{base}(i)}$: position of track i in the final ranked list generated in baseline; $\boldsymbol{rscore_{personal}(i)}$: ranking score of track i in personalized method; $\boldsymbol{rscore_{base}(i)}$: ranking score of track i in baseline

1: $rscore_{personal}(i) \Leftarrow n - r_{personal}(i) + 1$
2: $rscore_{base}(i) \Leftarrow n - r_{base}(i) + 1$
3: **for** each track$i$ **do**
4: $\quad score_{mix}(i) \Leftarrow 0.5 * rscore_{personal}(i) + 0.5 * rscore_{base}(i)$
5: **end for**
6: return top 10 tracks with the highest $score_{mix}$ value

4 EXPERIMENT DESIGN

We conducted an online experiment to investigate whether a preference-based playlist generated by the personalized method or mixed method can better help the user to explore a new genre than a simple genre-typical playlist generated by the baseline method. We adopted a comparative design as used by Ekstrand et al. [7] to precisely measure the perceptual differences between the baseline method and one of the preference-based methods (see Figure 2). The comparative design allows participants to compare the perceived difference between the two presented playlists at the same time. During the experiment, they were asked to evaluate the perceived differences between the simple genre-typical playlist and one of two preference-based playlists: either a playlist that was strongly personalized (generated by the personalized method) or a playlist that was a mix of the genre-typical tracks and personalized items (generated by the mixed method). There was one between subjects-condition in the experiment: compare the genre-typical playlist with the pure personalized playlist or the genre-typical playlist with the mixed playlist.

Participants were invited by email and informed of the basic information about the study. Most of them were from the participant pool of Eindhoven University of Technology. All participants were required to have an active Spotify account. We raffled 15 euro among every 5 participants using a lottery (expected value = 3 euro per person).

The experiment started after participants agreed with the informed consent form and logged in with their Spotify account. Participants were then redirected to a page where they needed to fill in a Musical Sophistication Index survey [22] used to measure their music expertise with two indicators: *Active Engagement* and *Emotion*. After testing the headphones, they were asked to select a new genre to explore. Next, two playlists of 10 recommended tracks based on the two algorithms (randomized, either the genre-typical playlist with the pure personalized playlist or the genre-typical playlist with the mixed playlist) would be generated. We presented the two playlists as List A and List B (the order was also randomized). Participants could listen to the tracks as they like. During

Table 2: Questions from the user survey. All SEM factor loadings are significant $p < .001$. Items without a factor loading were excluded from the analysis

Considered aspects	Items	SEM Coef.
Accuracy	Which playlist has more songs that you find appealing?	0.949
Alpha: 0.96	Which playlist has more songs that you might listen to again?	0.942
AVE: 0.87	Which playlist has more obviously bad songs for you?	
	Which playlist has more songs that are well-chosen?	
Personalization (formerly)	Which playlist better understands your tastes in music?	0.933
	Which playlist seems more personalized to your music tastes?	0.876
	Which playlist has fewer songs you feel familiar with?	
	Which playlist has more songs with styles that you like to listen to?	0.947
Representativeness	Which playlist better represents the mainstream tastes of the genre?	
Alpha: 0.81	Which playlist has more songs matching the style of the genre?	0.818
AVE:0.65	Which playlist has fewer songs you would expect from the genre?	−0.772
	Which playlist seems less typical of the genre?	−0.779
Helpfulness	Which playlist better supports you to get to know the new genre?	0.716
Alpha: 0.77	Which playlist motivates you more to delve into the new genre?	
AVE: 0.61	Which playlist is more useful to explore a new genre?	0.626
	Which playlist has more songs that helps you understand the new genre?	0.402
Diversity	Which playlist has more songs that are similar to each other?	
Alpha: N.A.	Which playlist has a more varied selection of songs within the genre?	
AVE: N.A.	Which playlist would suit a broader set of tastes?	
	Which playlist has songs that match a wider variety of moods?	

the whole process, we recorded their interaction with the songs. They could also save the generated playlists directly to their Spotify account. We believe that a click on this button suggests the user is taking action towards the change and going beyond the contemplation stage as in the transtheoretical model. At the same time, they needed to fill in a survey which contained comparative questions on (1) perceived personalization, (2) perceived representativeness of the playlist for the genre selected, (3) perceived helpfulness to explore the new genre, (4) perceived accuracy of the tracks and (5) perceived diversity of the generated playlist. Each factor was measured by 4 questions. We adapted the questions from Ekstrand et al. [7] regarding accuracy and diversity and constructed new items for three additional factors: personalization, representativeness and helpfulness. Note that each question asked for the relative difference between the two playlists. The full list of questions can be found in Table 2.

5 RESULTS

The online experiment ran in January 2019 with 156 valid responses (we excluded 7 users who either finished the questionnaire too fast or had no interactions with the recommended items according to our log files). The average age of participants was 23.05 years ($SD = 5.48$), with 78 females and 78 males. Participants were randomly assigned to either a condition with the baseline against the personalized list (N=69) or the baseline against the mixed (N=87) list. Order of the list presentation in the display was randomized and did not result in any significant differences in our initial analyses, so we exclude the discussion of order from the rest of the results section. Table 3 shows the number of participants that selected each genre to explore. The generated playlists are sufficiently different in both conditions. In the baseline against personalized list condition, only 3 out of 69 users got some duplicated tracks. In the baseline against the mixed condition, 42 out of 87 users in the mixed condition got duplicated tracks. However, the mean number of the duplicated tracks is 2.3 ($std = 1.3$, from the 42 users with duplicated tracks), which is small considering it is a mix of the genre-typical list and personalized list. Besides, the median number is 2 and only 3 users got the max of 5 duplicated tracks.

5.1 Structural Equational Model

The main purpose of the study is to find out which type of genre exploration is more helpful, and how perceptions of accuracy, personalization, diversity and representativeness of the list can explain why that type is more helpful. To test these relations, we subjected the survey responses to confirmatory factor analysis (CFA) and structural equation modeling (SEM), following the user-centric framework by Knijnenburg et al. [15, 16].

The CFA treated all question responses as ordinal. We coded all questions such that higher scores mean less for the baseline and more for the other preference-based list. The CFA confirmed the factor structure as we designed it, but some questions were excluded due to low explanatory power (low R^2) or high cross loadings. The items of personalization and accuracy showed strong correlations and cross loadings, we therefore collapsed these into one factor accuracy. We then submitted these factors to a saturated SEM model, testing all possible structural relations and iteratively removing those that were not significant.

Figure 3 shows the final SEM model and Table 2 shows the factors and factor loadings of the factors included in the model. The model had a good fit ($\chi^2(81) = 107.076, p = .028, CFI = .997, TLI = .996, RMSEA = .045, 90\%\ CI : [0.016, 0.067]$). While constructing the SEM model, we dropped the diversity factor for reasons of

parsimony. Though diversity was related to accuracy and representativeness, it did not relate to any of the independent measures or helpfulness and therefore did not have much explanatory power or theoretical relevance. In building the model, we related the subjective constructs to the independent measure (whether the preference-based list was personalized or mixed), and checked how the effects were moderated by musical sophistication, as per hypothesis 5. We found that the Active Engagement score (MSAE) were correlated to some concepts in our model, and therefore included a dummy variable to indicate if this subject is high or low on MSAE, based on a median split.

Looking at the final SEM model, we observe that helpfulness is indeed positively related to both representativeness and accuracy, consistent with hypothesis 1. This means (in our relative comparison measurement) that whenever a list is perceived to be more accurate or representative than the other list, it is perceived more helpful in exploring the genre. The SEM model also shows that in general, users perceive mixed lists to be more representative of the genre than personalized lists (again both relative to the baseline list), as indicated by the positive effect of mixed on representativeness, consistent with hypothesis 2. Users high on MSAE tend to report somewhat higher accuracy (though not significant at p=.05), which means they provide somewhat higher scores for the preference-based lists (both mixed and personalized) than the baseline. Finally, we observe an interesting positive interaction between the mixed method and MSAE on helpfulness, which indicates that especially those high on MSAE tend to think the mixed method scores better on helpfulness than the personalized method.

We also tested the total indirect effect of the condition (mixed versus personalized) on helpfulness. Whereas there is no direct effect of mixed on helpfulness, the overall indirect effect of mixed (versus personalized) on helpfulness via representativeness is significant ($coef. = 0.276$, $se = 0.103$, $p < .01$) showing that representativeness mediates the effect: the mixed list is perceived to be more helpful than the personalized list because of the higher representativeness.

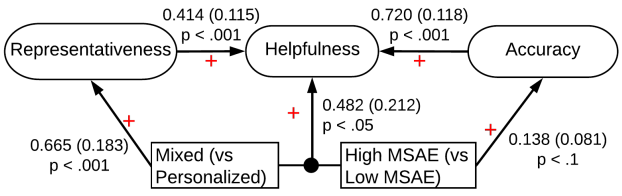

Figure 3: Structural Equation Model showing the relations between the measured constructs and the independent variables. Arrows represent the standardized coefficients with standard error between brackets and p-values.

5.2 Absolute differences between methods

Though the SEM model provides valuable insights in the relations between the subjective constructs, it has a strong limitation: in building the factors, the bipolar nature of our response scale gets lost, and we can only compare relatively, as is clear from the results discussed in the previous section. The bipolar scale with midpoint however allows us to also make absolute statements of the type: "The baseline is more representative than the mixed method", which is required to answer some of our hypotheses and get deeper insights in what method might be more helpful.

We took the questions that were found to be good predictors in the CFA/SEM for each concept and average them into an overall score ranging from -2 (baseline method is much more than other) to 2 (preference-based method is much more than baseline), with the zero midpoint to test against, using one-sample t-tests. In this analysis, we also included diversity for completeness, even though it was not part of the SEM.

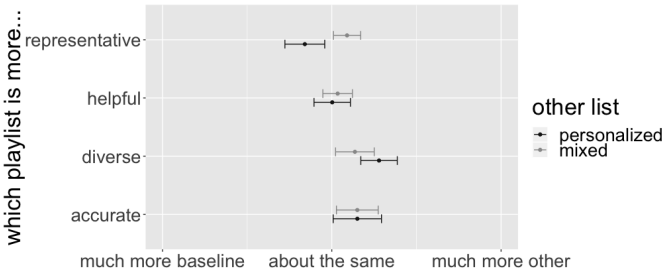

Figure 4: Perceived absolute difference for each subjective construct comparing the baseline to the other (preference-based) list for each concept across the two conditions. The error bars represent the 95% confidence interval.

Figure 4 shows that for all subjective constructs, the mean and 95% confidence intervals fall outside of the midpoint, except for helpfulness. For both mixed ($t(86) = .78, p = .43$) and personalized ($t(68) = .045, p = .96$) conditions, helpfulness scores are not different from the zero midpoint, suggesting that the baseline is not more or less helpful than the preference-based methods.

Looking at representativeness, the personalized condition is significantly below the midpoint ($t(68) = -2.70, p = 0.008$), suggesting that the baseline is perceived to be more representative than the personalized list. This supports our hypothesis 2. The opposite holds for the mixed condition, which is perceived to be even more representative than the baseline ($t(86) = 2.22, p = .03$). This is inconsistent with H2 in which we hypothesized that the baseline would always be more representative then any of the preference-based methods. A possible reason might be that users can better recognize representative items when the items also match their tastes.

Consistent with hypothesis 3, users perceived both the recommendations from the mixed method ($t(86) = 2.22, p = .03$) and the personalized method ($t(68) = 2.11, p = .04$) to be more accurate than the baseline. This is expected since both methods are generated with items from the preference-based method. Consistent with hypothesis 4, users perceived the playlists from the mixed method to be more diverse than the baseline ($t(86) = 2.38, p = .02$). Moreover, the perceived diversity of the playlist from the personalized method is also higher than the baseline ($t(68) = 5.17, p < .001$). Though we did not hypothesize any difference, a possible reason could be

Table 3: Number of participants that selected this genre to explore

Genre	avant-garde	blues	classical	country	electronic	folk	jazz	new-age	rap	rnb
N (participants)	10	17	11	9	34	11	26	13	15	10

there are many different ways to balance the user's preference on different audio features when the genre is away from the user's current preference, thus make the recommendations diverse. On the other hand, recommendations from the baseline method are always of genre typical tastes and therefore could be perceived less diverse, which is clearly suggested by our data. However, as we noted in the SEM model, diversity was not related to helpfulness nor influenced by our conditions.

5.3 Which is more helpful?

Both our SEM model as well as the analysis of the absolute differences are inconclusive as to what approach is the most helpful for users to explore a new genre. The SEM model does show an indirect effect of mixed, as well as an interaction of the mixed condition with users' MSAE scores (based on a median split). However, that interaction only reflects the relative differences between the conditions. Looking again at the absolute scores on the scale, we might be able to better understand these effects. We plot this interaction, including the 95% confidence intervals in Figure 5. Looking at the figure, we observe that for users with low MSAE for both preference-based methods the means are around the midpoint: in these conditions the baseline is not regarded to be more helpful than either of the preference-based methods (all t-tests are non-significant). For users high on MSAE, we observe stronger differences, with the mean of the those in the mixed condition significantly above the midpoint ($t(35) = 2.24, p = .03$), showing that users high on MSAE experience the mixed playlist more helpful than the baseline. The mean of the personalized list for high MSAE users is lower than the midpoint, but again not significantly.

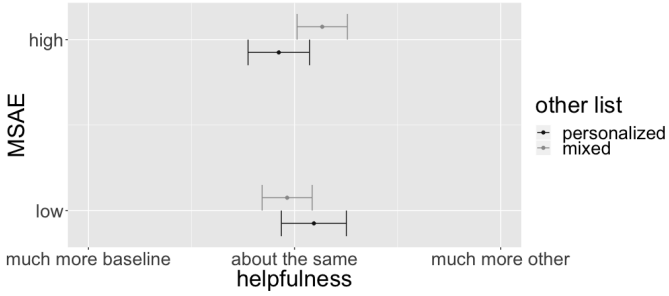

Figure 5: Perceived absolute difference on helpfulness comparing the baseline to the other (preference-based) list for users with low/high MSAE across the two conditions. The error bars represent the 95% confidence interval.

5.4 Behavioral data

Apart from user perceptions, we also measured user interaction in the interface by checking which songs users interacted with, but did not find any differences between conditions or between high or low MSAE scores. We also allowed users to save either of the two playlists or both to their Spotify accounts, by means of a button below the lists. Between conditions and other factors, we find that 15% to 22% of the users have saved either playlist to their Spotify account. For example, 19% of the preference-based playlists were saved to their accounts, against 16.7% of the baseline playlists. People high on MSAE tend to save playlists more (21.9%) than people low on MSAE (14.4%). However, none of these differences are significant. Therefore, we cannot draw any conclusions on which type of playlists convinces users to take more action to further explore the new genre in the future. A larger sample size would be required than we currently have to draw conclusions on these behavioral measures.

6 CONCLUSION AND DISCUSSION

To conclude, users perceived no difference in helpfulness across the conditions, so we cannot state concludingly whether a preference-based playlist or a genre-typical playlist can better help the user to explore the new genre. However, our SEM model shows that the perceived helpfulness is positively related to both perceived accuracy and representativeness. A more personalized playlist could be far from the genre-typical tastes and more close to the user's current taste, thus makes it easier to take steps towards change. While a more representative (genre-typical) playlist could be further away from the user's current taste but can better help the user understand the typical taste or characteristics of the new genre. Therefore, it is important to balance the two conflicting factors to increase the perceived helpfulness of the playlist as our mixed method does. Although users did not perceive the mixed method to be more helpful in general, we do find that users with high MSAE perceived the mixed method to be more helpful than the purely personalized method. This indicates that the mixed method could be a good way to balance the perceived accuracy and representativeness to increase the perceived helpfulness for new genre exploration.

There are also some limitations in our work. We were intended to use the behavior data from the 'save' action to understand more about users' choice, however, not many people saved the playlists (the reason might be the button was placed at the bottom of the webpage and might have been ignored) and some people saved both playlists. For our future work, we should ask about user's choice by directly asking questions, such as "Which playlist would you like to save to your Spotify account?" as in the previous work [7]. Besides, the current mixed method set the weight to be fixed (50%). However, we might want to explore more on the influence of weight, for example, it might be different for people with high or low MSAE. Furthermore, our current interface does not support any interactive exploration from the user side. For future work, it could be interesting if users can also guide their exploration with the help of the system, combining our genre exploration with existing exploration tools.

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*. ACM, 275–279.

[2] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 119–126.

[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[4] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. 2013. Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management* 49, 1 (2013), 13–33.

[5] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2013. LinkedVis: exploring social and semantic career recommendations. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 107–116.

[6] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. 2015. Cross-domain recommender systems. In *Recommender Systems Handbook*. Springer, 919–959.

[7] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. 2014. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 161–168.

[8] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction* 4, 2 (2011), 81–173.

[9] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 221–224.

[10] Mark P Graus and Martijn C Willemsen. 2015. Improving the user experience during cold start through choice-based preference elicitation. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 273–276.

[11] F Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting users in control of their recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 3–10.

[12] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[13] Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A Konstan, and Paul Schrater. 2015. I like to explore sometimes: adapting to dynamic user novelty preferences. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 19–26.

[14] Komal Kapoor, Nisheeth Srivastava, Jaideep Srivastava, and Paul Schrater. 2013. Measuring spontaneous devaluations in user preferences. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1061–1069.

[15] Bart P Knijnenburg and Martijn C Willemsen. 2015. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook*. Springer, 309–352.

[16] 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 (2012), 441–504.

[17] Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. 2015. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 975–984.

[18] Benedikt Loepp, Tim Hussein, and Jüergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3085–3094.

[19] Judith Masthoff. 2004. Group modeling: Selecting a sequence of television items to suit a group of viewers. In *Personalized digital television*. Springer, 93–141.

[20] Judith Masthoff. 2015. Group recommender systems: aggregation, satisfaction and group attributes. In *recommender systems handbook*. Springer, 743–776.

[21] Leigh McAlister and Edgar Pessemier. 1982. Variety seeking behavior: An interdisciplinary review. *Journal of Consumer research* 9, 3 (1982), 311–322.

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

[23] Sayooran Nagulendra and Julita Vassileva. 2013. Providing awareness, understanding and control of personalized stream filtering in a p2p social network. In *International Conference on Collaboration and Technology*. Springer, 61–76.

[24] Sayooran Nagulendra and Julita Vassileva. 2014. Understanding and controlling the filter bubble through interactive visualization: a user study. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 107–115.

[25] T Nava, Shahin Rostami, and Barry Smyth. 2018. Knowing the unknown: visualising consumption blind-spots in recommender system. (2018).

[26] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[27] James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion* 12, 1 (1997), 38–48.

[28] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. 2015. Music recommender systems. In *Recommender Systems Handbook*. Springer, 453–492.

[29] Markus Schedl, Peter Knees, and Gerhard Widmer. 2006. Investigating web-based approaches to revealing prototypical music artists in genre taxonomies. In *Digital Information Management, 2006 1st International Conference on*. IEEE, 519–524.

[30] 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*. International World Wide Web Conferences Steering Committee, 987–997.

[31] 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*. ACM, 513–521.

[32] 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*. ACM, 1383–1392.

[33] 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*. ACM, 335–338.

[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*. ACM, 351–362.

[35] Martijn C Willemsen, Mark P Graus, and Bart P Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction* 26, 4 (2016), 347–389.