How playlist evaluation compares to track evaluations in music recommender systems

Sophia Hadash
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
s.hadash@tue.nl

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, The Netherlands
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
m.c.willemsen@tue.nl

ABSTRACT

Most recommendation evaluations in music domain are focused on algorithmic performance: how a recommendation algorithm could predict a user's liking of an individual track. However, individual track rating might not fully reflect the user's liking of the whole recommendation list. Previous work has shown that subjective measures such as perceived diversity and familiarity of the recommendations, as well as the peak-end effect can influence the user's overall (holistic) evaluation of the list. In this study, we investigate how individual track evaluation compares to holistic playlist evaluation in music recommender systems, especially how playlist attractiveness is related to individual track rating and other subjective measures (perceived diversity) or objective measures (objective familiarity, peak-end effect and occurrence of good recommendations in the list). We explore this relation using a within-subjects online user experiment, in which recommendations for each condition are generated by different algorithms. We found that individual track ratings can not fully predict playlist evaluations, as other factors such as perceived diversity and recommendation approaches can influence playlist attractiveness to a larger extent. In addition, inclusion of the highest and last track rating (peak-end) is equally good in predicting playlist attractiveness as the inclusion of all track evaluations. Our results imply that it is important to consider which evaluation metric to use when evaluating recommendation approaches.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; Heuristic evaluations; • **Information systems** → *Recommender systems*; *Relevance assessment*; *Personalization.*

KEYWORDS

User-centric evaluation, recommender systems, playlist and track evaluation

IntRS '19, September 19, 2019, Copenhagen, Denmark

ACM Reference Format:
Sophia Hadash, Yu Liang, and Martijn C. Willemsen. 2019. How playlist evaluation compares to track evaluations in music recommender systems. In *Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS '19).* CEUR-WS.org, 9 pages.

1 INTRODUCTION

In user-centric evaluation of personalized music recommendation, users are usually asked to indicate their degree of liking of individual tracks [2, 4, 5, 14] or by providing a holistic assessment of the entire playlist (e.g. playlist satisfaction or playlist attractiveness) [6, 9, 12, 16, 17] generated by the recommendation approaches. Most recommender evaluations are focused on the first type of evaluation to test algorithmic performance: can we accurately predict the liking of an individual track. Many user-centric studies in the field [15] however focus on the second metric, does the list of recommendations provide a satisfactory experience. Often these studies find playlist satisfaction is not just about the objective or subjective accuracy of the playlist, but also depends on the difficulty of choosing from the playlist or playlist diversity [27]. For example, in the music domain perceived diversity of the playlist has been shown to have a negative effect on overall playlist attractiveness [7]. Bollen et al. [3] showed people were just as satisfied with a list of 20 movie recommendations which included the top-5 list and a set of lower ranked items (twenty's item being the 1500th best rank) as with a list of the best 20 recommendations (top-20 ranked).

Research in psychology also shows that people's memory of overall experience is influenced by the largest peak and end of the experience rather than the average of the moment to moment experience [13]. Similar effects might occur when we ask users to evaluate holistically a list on attractiveness: they might be triggered more by particular items in the list (i.e. ones that they recognize as great (or bad), ones that are familiar rather than ones that are unknown, cf. mere exposure effect [25]) and therefore their overall impression might not simply be the mean of the individual ratings.

These results from earlier recommender research and from psychological research suggest that overall (holistic) playlist evaluation is not just reflected by the average of liking or rating of the individual items. However, to our best knowledge, no previous work has explored the relation between users' evaluation of individual tracks and overall playlist evaluation. To some extent this is because it is not common that both types of data are collected in the same study. Therefore, in this work, we would like to investigate how individual

item evaluations relate to holistic evaluations in sequential music recommender systems.

We explore these relations using a within-subject online experiment, in which users are asked to give individual ratings as well as overall perception of playlist attractiveness and diversity in three conditions: (1) track and artist similarity algorithm (*base*), (2) track and artist similarity algorithm combined with genre similarity algorithm (*genre*) and (3) track and artist similarity algorithm combined with audio feature algorithm (*gmm*). The track and artist similarity algorithm can be regarded as a low-spread strategy since recommendations are generated from a small subset of the total pool of tracks relatively close to the user's tastes [8]. Both the *genre* approach and *gmm* approach are high-spread strategies which generates user-track ratings for a large proportion of the total pool of tracks.

In this study, we are interested in how perceived attractiveness of the playlist is related to perceived playlist diversity and individual track ratings across the three conditions. In addition, we also include a set of objective features of the playlist in the analysis. We test whether that users' perceived attractiveness of the playlist will also be affected by (1) the peak-end effect: the track they like most and the end track, (2) their familiarity to the recommendations in the playlist and (3) occurrences of good recommendations in the playlist: people might be satisfied with a playlist as long as at least some recommendations are good.

2 RELATED WORK

2.1 User-centric evaluation in music recommendation

User-centric evaluation for recommendation approaches is necessary in order to understand users' perception of the given recommendations [15], such as acceptance or satisfaction [23, 24].

User-centic evaluation in music recommendation can be at individual track level or whole playlist level. Users' perception towards the whole playlists are often measured under the context of automatic playlist generation [20], smooth track transition [9] or when the goal is to evaluate the whole recommender system [12, 17]. For example, users were asked to indicate their perception towards the recommended playlists to investigate how different settings of control in the recommender system influence their cognitive load as well as their acceptance to the recommendations [12]. However, when it comes to the evaluation of the recommendation algorithms, users are often asked to indicate their ratings[2, 4] for each individual track rather than the playlist as a whole, neglecting the fact that tracks are often listened in succession or within a playlist.

Individual item ratings can not fully reflect users' degree of liking towards the recommendation list. Perceived diversity is a factor that can only be measured at the list level. Willemsen et al. [27] has shown that perceived diversity of a movie recommendation list has a positive effect on perceived list attractiveness and a higher perceived diversity would make it easier for users to make a choice from the recommendations. Ekstrand et al. [6] also show that perceived diversity has a positive effect on user satisfaction. While in music domain, Ferwerda et al. [7] found that perceived diversity has a negative effect on perceived attractiveness of the recommendation list, however, this effect turns to positive when the recommendation list can help users to discover new music and enrich their music tastes.

The novel contribution of this work is that we include both measurements in the study for personalized music recommendations, aiming to uncover the relation between individual track evaluation and holistic evaluation of music playlists.

2.2 Peak-end effect

Research in psychology has looked into the differences between the 'remembering self' and the 'experiencing self' [13], as reflected in the peak-end rule: the memory of the overall experience of a painful medical procedure is not simply the sum or average of the moment to moment experience, but the average of the largest peak and the end of the experience.

In music domain, several studies have found that the remembered intensity of the music listening experience is highly correlated with peak, peak-end and average moment-to-moment experience [21, 22]. However, it is argued by Wiechert [26] that these studies fail to consider users' personal musical preferences and that the peak-end value and the average value measured in the studies might be correlated with each other. Rather than giving participants the same stimuli, Wiechert gave participants a list of songs based on their current musical preference and came up with a new metric: pure peak-end value (the difference between peak-end and average). He found that while the average experience could explain a significant part of playlist experience variance, the pure peak-end value could explain a part of variance that would not be explained by the average.

3 METHOD

In this study three algorithms are used for generating playlists. These algorithms are designed to use user preferences in the form of (ordered) lists of tracks, artists, or genres a user is known to like. The advantage of using such an input form is that these algorithms can be used with user preferences obtained from commercial platforms. In this study Spotify [1] user profiles are used. These preferences are in the form of ordered lists of top tracks and artists. The first algorithm is based on track and artist similarity. The second algorithm uses a genre similarity metric based on genre co-occurrence among artists. The third algorithm recommends tracks based on a Gaussian mixture model on track features derived from audio analyses (see [10] for details). All algorithms are described in detail in [8].

3.1 Track and artist similarity algorithm

The track and artist similarity algorithm is a combination of the same sub-algorithm applied to both a list of tracks and artists a user is known to like. The input to this sub-algorithm is a list of items, potentially ordered on user likeability. This sub-algorithm uses Spotify's seed recommendation system to explore items that are similar to the input. Based on the occurrence of items in the results, an output list is generated with user likeability prediction scores. The algorithm is formulated in Algorithm 1, with an illustration by example in Figure 1.

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

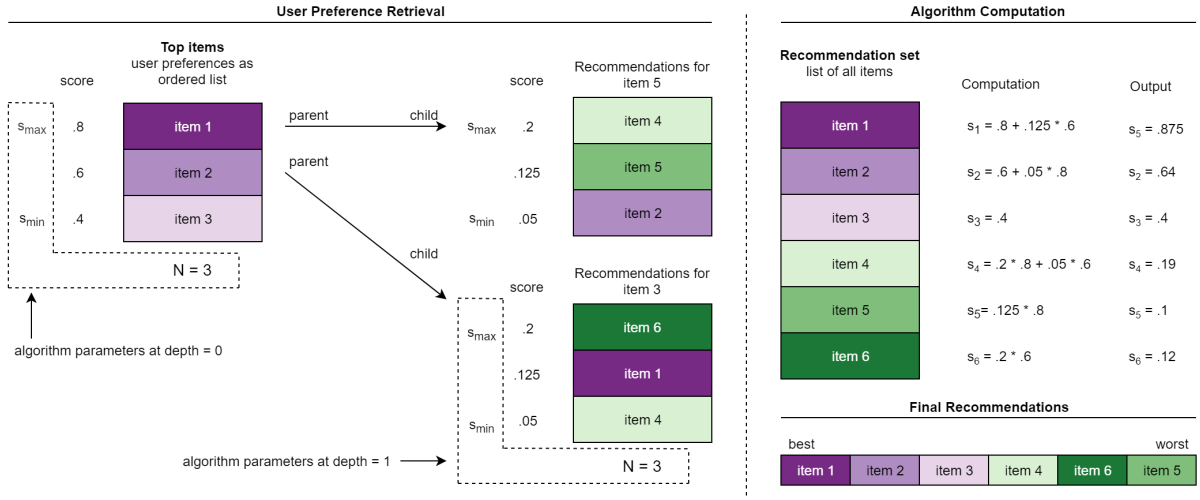

Figure 1: Illustration of the track and artist similarity algorithm using an example.

Algorithm 1 Track and artist similarity

s_i: score of item i, $\boldsymbol{x}$: temporary score of item i, $\boldsymbol{recs}$: recommendation set, $\boldsymbol{N}$: number of sibling nodes, $\boldsymbol{pos_{node}}$: position of the node in its parents' children, $\boldsymbol{s_{min}}, \boldsymbol{s_{max}}$: scores assigned to the last and first sibling node at the current tree depth.

```
for each item i in recs do
    s_i = 0
    for each node_j as current_node where node_j is item i do
        x = current_node.score
        while current_node.parent not null do
            current_node = current_node.parent
            x = x * current_node.score
        end while
        s_i = s_i + x
    end for
end for
return (recs, s) order by s descending

def node.score:
    return (N - pos_node)/N * (s_max - s_min) + s_min
```

3.2 Genre similarity algorithm

The genre similarity algorithm uses an ordered list of genres the user likes S'_u (a column vector which shows the user degree of liking to all genres built from the user's top artists) and a similarity metric $\boldsymbol{D}$ to generate genre likeability scores for other genres. Then, the resulting extrapolated list $\boldsymbol{S_u}$ is used to favor recommendations from genres with high likeability scores.

There are 1757 different types of genres available in our dataset, therefore both S'_u and S_u are column vectors of dimension 1757 and matrix $\boldsymbol{D}$ is of dimension 1757×1757.

The similarity metric is based on co-occurrence analysis of artists, similar to the methodology used in [19]. The co-occurrence analysis used a database consisting of $n \approx 80.000$ artists. For each artist it was known which genres he/she produced music in. The data is extracted from Spotify's developer API. The co-occurence analysis generated a normalized symmetric similarity matrix $\boldsymbol{D}$. The likeability scores of the user towards the list of genre is then computed as follows, where $\boldsymbol{I}$ is the identity matrix.

$$S_u = (D + I)S'_u \tag{1}$$

3.3 Audio feature algorithm

The audio feature algorithm clusters tracks with similar audio features using a Gaussian mixture model (GMM). A database of $n \approx 500.000$ tracks containing 11 audio analysis features were used to train the model. The audio features consisted of measures for danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Multiple GMM's were fitted using the expectation-maximization (EM) algorithm for varying component numbers. The model with 21 components had the lowest BIC criterion and was therefore selected. Then, cluster likeability was computed as follows (see [8]):

$$p(user\ likes\ cluster\ i) = \frac{1}{N_{top}} \sum_{j=1}^{N_{top}} p(track\ j\ belongs\ to\ cluster\ i) \tag{2}$$

Finally, the output recommendations favored tracks corresponding to clusters with high user likeability probabilities.

3.4 Familiarity of the recommendations to the users

Both the track and artist similarity and the genre similarity algorithms generate recommendations close to the users' known preferences. Recommendations are based on artists and genres that are familiar to the user. The audio feature algorithm on the other hand recommends tracks based on audio feature similarity. As a result, recommended tracks are more likely to have genres and artists that are less familiar to the users.

4 EXPERIMENTAL DESIGN

To evaluate the relation between track evaluations and playlist evaluations, a within-subjects online experiment was conducted. The study included three conditions in randomized order: track and artist algorithm (*base*), track and artist algorithm combined with the genre similarity algorithm (*genre*), and track and artist algorithm combined with the audio feature algorithm (*gmm*). In each condition participants were presented with a playlist containing 10 tracks generated by the corresponding algorithm and evaluated the individual tracks on likeability and personalization and the playlist as a whole on attractiveness and diversity. The playlist included the top 3 recommendations and the 20^{th}, 40^{th}, 60^{th}, 80^{th}, 100^{th}, 200^{th}, and 300^{th} recommendation in random order. Lower ranked recommendations were included such that algorithm performance could be evaluated more easily, as lower ranked recommendations should result in lower user evaluations.

4.1 Participant Recruitment

Participants were primarily recruited using the JF Schouten participant database of Eindhoven University of Technology. Some participants were recruited by invitation. Participants were required to have a Spotify account (free or Premium) and to have used this account prior to taking part in the study.

4.2 Materials

The track evaluations included likeability and personalization measures. One question was used for each of the tracks. This was decided based on the repetitive nature of individual track evaluations. The questions for measuring track likeability was: "Rate how much you like the song". For measuring perceived track personalization we used the following item: "Rate how well the song fits your personal music preferences". Both questions were answered on a 5-point visual scale with halves (thus 10 actual options) containing stars and heart icons as shown in Figure 2.

The playlist evaluation included playlist attractiveness and playlist diversity and is presented in Table 1.

Additional scales used in the study were a demographics scale and the Goldsmith Music Sophistication Index (MSI) [18]. The demographics scale measured gender, age, and Spotify usage. Spotify usage was measured using a single item: "I listen to Spotify for __ hours a week" with 7 range options.

Table 1: The playlist evaluation scale

Concept	Item
Perceived attractiveness Alpha: .94	The playlist was attractive
	The playlist showed too many bad items
	The playlist matched my preferences
Perceived diversity Alpha: .85	The playlist was varied
	The tracks differed a lot from each other on different aspects
	All the tracks were similar to each other

Note. The scale is adapted from [27].

4.3 Study Procedure

After consenting, participants were prompted with a login screen where they could connect their Spotify account with the study. Participants who did not have a Spotify account or who had a Spotify account containing no user preference data could not continue with the study. After Spotify login, participants completed a background survey. In the survey they reported their Spotify usage and music sophistication.

Following the background survey, the user entered the track evaluation phase in which a playlist was presented to the user generated by one of the algorithms. The interface (see Figure 2) contained an interactive panel showing the tracks of the playlist, a survey panel in which they had to rate the tracks, and a music control bar. Participants could freely browse through the playlist while providing the ratings. After all ratings were provided, participants entered the playlist evaluation phase in which they answered the playlist evaluations questions (Table 1). The track evaluation phase and playlist evaluation phase were then repeated for the remaining conditions.

Finally, participants were thanked for their time and were entered into a reward raffle. Among every 5 participants one participant received 15 euro compensation. In total the study lasted approximately 15 minutes.

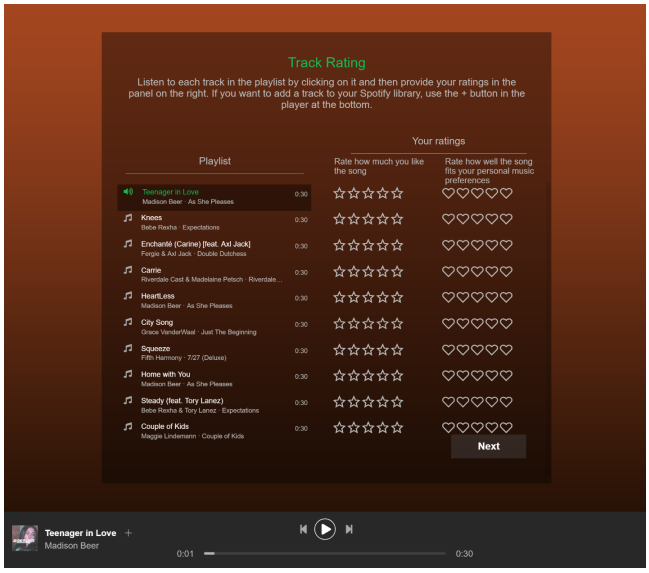

Figure 2: Preview of the track rating screen as displayed to the participants during the study.

5 RESULTS

Participants in this study included 59 people, of which 54 were recruited through the JF Schouten database. The sample consisted of 31 males and 28 females. The age of the participants ranged from 19 to 64 ($M = 25.6, SD = 8.8$). On average participants listened to Spotify for 7 to 10 hours per week. MSI scores ranged between 0 and 5 ($M = 2.18, SD = 1.0$). The study took place between 9^{th} of January and 1^{st} of February of 2019.

We found that there was no effect of personalization rating on perceived attractiveness, while likability rating can partially predict perceived attractiveness. Furthermore, playlist attractiveness was more strongly related to the recommendation algorithm. Playlists in the *gmm* condition were less positively evaluated compared to playlists in the other conditions even though the track evaluations were similar on average. In other words, while participants evaluated tracks across conditions similarly, the playlist evaluations differed substantially (see Figure 3).

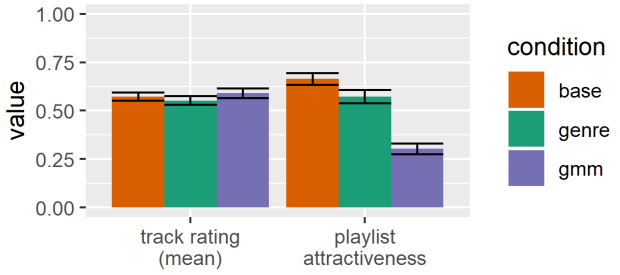

Figure 3: Participants' subjective evaluations of individual tracks (left) and playlists (right). The error bars indicate the standard error.

5.1 Overview of statistical methods

The results are analyzed using three methodologies. The first methodology concerns the performance of the recommendation algorithms. This was analyzed using descriptive statistics concerning the relation between recommendation scores predicted by the algorithms and the user ratings.

In the second methodology the relation between playlist evaluations and track ratings was aggregated on the playlist-level (i.e. 3 observations per user). In this methodology, an aggregate measure for track evaluation was used, more specifically, three aggregation measures: mean (Model 1), peak-end (Model 2), occurrence of at least a 3-star rating (Model 3). Using these aggregates, a linear mixed-effects model was used such that variation in participants' answering style can be included as a random-effects variable. Playlist diversity and the recommendation approaches were included as fixed-effects variables in Model 1a, Model 2a and Model 3a, and interaction-effects were included in Model 1b, Model 2b and Model 3b.

Finally, the last methodology explores how variations within the track-level may explain playlist attractiveness. This analysis used a linear mixed-effects model on the track level (i.e. 3x10 observations per user) (see Table 3: Model 4) with participants modelled as a random-effects variable, similar to the playlist-level analysis. For the track-level variables four types of indicators were included additional to the rating, condition, and diversity. The first indicator indicates whether the track was high-ranked (top 3 recommendation) or low-ranked (top 20 to 300). The second indicates for each track whether it was the highest rating of the playlist. Thus, if a user gave two 4-star ratings and 8 lower ratings, the variable would indicate those two tracks with a 1, otherwise 0. The third indicator is the familiarity which shows whether a track was predicted to be familiar to the user based on their top tracks and artists. Finally, the last indicator contains the playlist order. This variable indicates whether the track was among the list which the user evaluated firstly, secondly, or thirdly.

5.2 Model results

5.2.1 Algorithm performance. The relation between recommendation scores and user evaluations of tracks is depicted in Figure 4. The illustration indicates that differences exist between algorithms in their performance on track evaluations. This is supported by an analysis of variance (ANOVA), $F(2, 171) = 36.8$, $p < .001$. The graph shows that for all algorithms, higher recommendation scores result in higher user ratings, showing that indeed tracks that are predicted to be liked better also get higher ratings. However, consistent with Figure 3, the scores for the *base* condition are consistently higher than for the other two algorithms. For the *genre* condition the slope seems to be steeper than for the other two conditions, showing that in this condition, user ratings are more sensitive to the predicted recommendation scores.

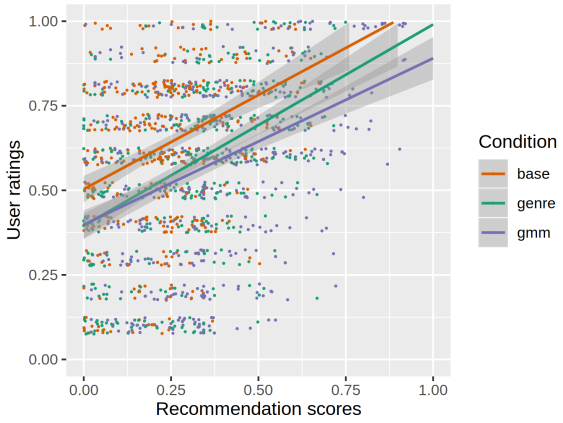

Figure 4: The relation between subjective user ratings and recommendation scores predicted by the algorithms. The user ratings are slightly jittered for the scatterplot only. The shaded area represents the 95% confidence interval.

5.2.2 Playlist-level relation between track evaluations and playlist evaluations. In this analysis, the effect of track evaluations on playlist evaluations is explored on a playlist-level, using three different aggregation measures (Models 1-3).

The effect of track evaluations on playlist attractiveness is illustrated in Figure 5. All three aggregation measures are very similar in predicting playlist attractiveness (see Table 2). We see a positive effect of the aggregating measure, indicating that if a user scores higher on that measure, she also finds the playlist more attractive, together with negative effects of the conditions *genre* and *gmm* consistent with the effect in Figure 3 that *gmm* and *genre* score lower than the *base* condition. The aggregate indicates occurrence of at least a 3-star rating (model 3) is a slightly worse predictor

Table 2: Playlist attractiveness by aggregated track evaluations (playlist-level)

	Mean		Peak-end		Positive	
	Model 1a	**Model 1b**	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**
rating (aggregate)	0.319^{***}	0.071	0.274^{**}	0.098	0.104^{*}	0.022
	(0.095)	(0.165)	(0.091)	(0.151)	(0.043)	(0.071)
genre	-0.090^{*}	-0.665^{***}	-0.081^{*}	-0.643^{***}	-0.095^{*}	-0.503^{***}
	(0.039)	(0.174)	(0.039)	(0.188)	(0.038)	(0.139)
gmm	-0.364^{***}	-0.741^{***}	-0.351^{***}	-0.840^{***}	-0.356^{***}	-0.730^{***}
	(0.039)	(0.175)	(0.038)	(0.194)	(0.038)	(0.129)
diversity	−0.059	-0.416^{**}	−0.067	-0.424^{**}	−0.078	-0.419^{**}
	(0.074)	(0.133)	(0.074)	(0.132)	(0.074)	(0.134)
rating (aggregate):genre		0.581^{*}		0.422^{*}		0.162
		(0.228)		(0.215)		(0.101)
rating (aggregate):gmm		0.127		0.230		0.118
		(0.224)		(0.217)		(0.101)
genre:diversity		0.428^{*}		0.444^{*}		0.481^{**}
		(0.183)		(0.183)		(0.185)
gmm:diversity		0.526^{**}		0.546^{**}		0.475^{**}
		(0.174)		(0.174)		(0.178)
Constant	0.512^{***}	0.865^{***}	0.492^{***}	0.836^{***}	0.620^{***}	0.889^{***}
	(0.076)	(0.135)	(0.086)	(0.141)	(0.062)	(0.102)
N	176	176	176	176	176	176
Log Likelihood	17.052	24.794	17.007	23.711	15.602	21.237
AIC	−20.105	−27.588	−20.013	−25.421	−17.204	−20.475
BIC	2.089	7.287	2.180	9.454	4.990	14.401
$R^2_{GLMM(m)}$	.351	.409	.342	.401	.330	.383
Random Effect						
# of Participants	59	59	59	59	59	59
Participant *SD*	0.063	0.053	0.08	0.054	0.083	0.064

Note. SD = standard deviation. The models are grouped by the method used for aggregating track evaluations. 'Mean' = mean value, 'peak-end' = average of highest rating and the last rating, 'positive' = indicator for occurrence of at least a 3-star evaluation. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

for playlist attractiveness compared to the mean and peak-end measures.

When the interaction-effects are included, the main-effect of ratings is no longer significant (models 1b, 2b and 3b) but we get several interactions of ratings with condition and condition with diversity. The interaction-effects of condition with perceived diversity and track evaluations are visualized in Figure 6 by separating the resulting effects by condition and we will discuss each condition and it's interactions separately.

The track evaluations had no effect on playlist evaluation in the *base* condition (they do for the other two conditions, as we will see below). Moreover, in the *base* condition, perceived diversity has a negative effect, indicating that playlists with high perceived diversity were less attractive compared to playlists with low perceived diversity. One potential explanation could be that since these playlists were constructed using a low-spread approach the recommendations were closely related to the users' known preferences (i.e. their top tracks that feed our algorithms). Therefore, the diversity in these users' preferences may have influenced the diversity of the recommended playlist. For instance, a person may listen to different genres during varying activities like working and sporting. The recommendations could then include music based on all these genres. While all recommendations are then closely related to the users' preferences and could receive potentially high evaluations, the playlist may not be very attractive due to the diversity in the genres.

In the *genre* condition, perceived diversity had no effect on playlist attractiveness. In this condition track evaluations strongly predicted playlist attractiveness regardless of diversity. The results show that though the *genre* playlist on average get a lower attractiveness score than the *base*, this effect is reduced when the aggregate ratings of the list are higher: in other words, only if users like the *genre* tracks, they like the playlist as much as the *base* one that has more low-spread, familiar tracks.

The *gmm* condition had similar results as the *genre* condition. Perceived diversity predicted attractiveness only marginally. However, while the track evaluations strongly predict attractiveness in

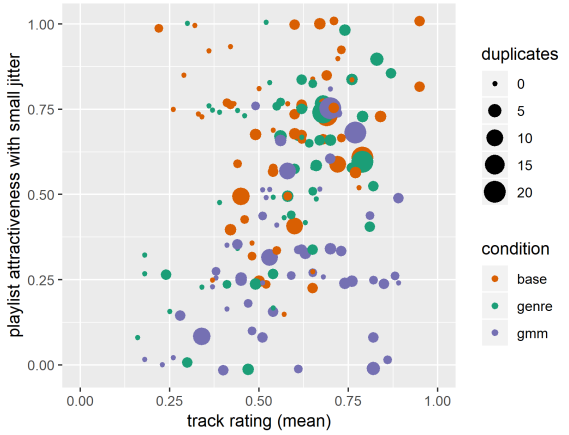

Figure 5: Playlist attractiveness by track rating (mean). The dot size indicates the number of duplicate items of the playlist in the playlists of the other conditions.

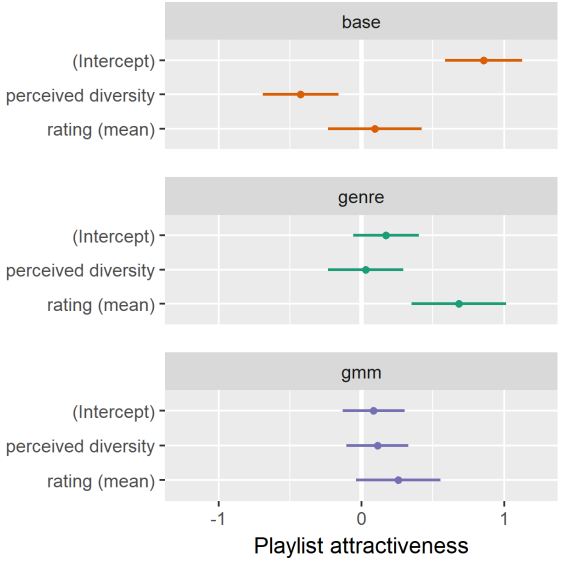

Figure 6: Linear model of playlist attractiveness by track ratings and condition for each condition.

the *genre* condition, it is only a weak predictor in the *gmm* condition. In other words, high aggregate ratings cannot really make up for the fact that the *gmm* list in general is evaluated worse than the *base* list. As in the *genre* condition this recommendation algorithm uses a high-spread approach and includes novel track recommendations. However, the *gmm* recommended tracks based on audio feature similarity is in contrast to genre similarity. Regardless of diversity or individual track evaluations, playlists using this approach were less attractive to participants.

Overall we find that overall attractiveness of a playlist is not always directly related to the liking of the individual tracks, as reflected by the aggregate ratings of the tracks, whether this is the mean rating, the peak-end value or the fact that at least one track is highly rated. We see that some conditions are more sensitive to these aggregate rating (*genre*) than the others. We also see an important (negative) role of diversity for the *base* condition in predicting overall attractiveness, but no effect in the other two conditions. In other words, different aspects affect playlist evaluation as recognized in the literature, but this highly depends on the nature of the underlying algorithm generating the recommendations.

Table 3: Playlist attractiveness by track evaluations (track-level)

	Model 4
rating	−0.009
	(0.020)
genre	-0.095^{***}
	(0.009)
gmm	-0.352^{***}
	(0.009)
diversity	-0.027^{***}
	(0.006)
high-ranked	0.003
	(0.009)
highest rating	0.002
	(0.012)
familiar	−0.011
	(0.012)
playlist order	0.012^{*}
	(0.005)
Constant	0.704^{***}
	(0.029)
N	1850
Log Likelihood	630.272
AIC	−1238.544
BIC	−1177.791
$R^2_{GLMM(m)}$	.307
Random Effect	
# of Participants	58
Participant *SD*	0.156

Note. SD = standard deviation, 'High-ranked' indicates the track was one of the top-3 recommendations, 'highest rating' indicates the track received the highest rating within that playlist for the participant, 'familiar' indicates whether the track was known to be familiar to the participant, 'playlist order' indicates whether the playlist was the first (=1), second (=2), or third (=3) list that the participant evaluated. Interaction terms as in Models 1-3 were omitted due to similarity to these models. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

5.2.3 Track-level relation between track evaluations and playlist evaluations. In this analysis, the effect of track evaluations on playlist evaluations is explored at track-level, trying to predict the overall attractiveness of each list with the individual track ratings, rather than the aggregate ratings. The results are shown in Table 3. Four

types of track-level variables are included in the analysis as described in Section 5.1.

Whether a track is high ranked or received the highest rating shows no significant effect on perceived attractiveness of the playlist. The track-level objective familiarity measures if the user is familiar with the artists of a track. The user is familiar with a track if at least one artist of the track also appears in the user's top listened tracks related artists. Although we expected there would be a positive effect of familiarity on playlist attractiveness (as also shown in [7]), there was no significant effect observed in model 4. A possible reason could be the objective familiarity measure was not sufficient to cover all tracks that the user is familiar with since it is only measured with the user's top tracks (the number is at most 50 for each user). In our future work, we are planning to directly ask for (self-reported) familiarity, rather than calculating these from the data. We also calculated a familiarity score for each track (how much the user is familiar with the track). We found that there was a positive correlation between objective familiarity and track ratings ($r_s(1770) = 0.326, p < .001$): users give higher ratings to tracks they are more familiar with, which is in line with previous work on mere exposure effect [1].

Playlist order is also a weak predictor of playlist attractiveness. Participants perceive the last playlist as the most attractive and the first as the least attractive. However, when interaction terms as in models 1-3 are included the effect is no longer significant. We also checked the condition orders generated by the random generator and found that each condition order occurred approximately equally often. In other words, the effect of condition order can not explain difference across conditions.

6 DISCUSSION OF RESULTS

We found that participants evaluate playlists on more aspects than merely the likeability of its tracks. Even though the tracks in recommended playlists may be accurate and receive positive user evaluations, playlists can still be evaluated negatively. In particular, the recommendation approach itself plays a role in the overall perceived playlist attractiveness.

One explanation may be that users have different distinct musical styles. Playlists that contain music from more than one of the users' styles may be less attractive to the user even though the track recommendations are accurate. Playlists in the *base* condition are most attractive, but suffer most from diversity. Users with multiple musical styles may have received playlists with music from multiple styles which could have been reflected in the perceived diversity of the playlist. Playlists from the *genre* condition were also based on genre similarity, in addition to the track and artist similarity. Therefore, if multiple musical styles are present in the user preferences, it is more likely in the *genre* condition that the musical style with the highest overall contribution overrules the music from the other musical styles. Furthermore, the *gmm* condition is least attractive. The recommendation algorithm used in this condition is based on audio feature similarity. Although tracks recommended in this condition were similar to the user preferences based on the audio features, they could be dissimilar based on more comprehensible attributes like genre and artists. It is likely that music from multiple musical styles were present in these playlists.

Another explanation may be the methodology of evaluation. While tracks are evaluated at the moment they are experienced, playlist evaluation occurs only after the tracks are experienced. Therefore, playlist evaluations are based on what users remember from the list. This difference may lead to differences in user evaluation styles. Although this may explain why differences occur between track and playlist evaluations, it cannot explain why the different recommendation approaches lead to different playlist attractiveness evaluations. Furthermore, using this explanation we would have expected a model improvement from the inclusion of the peak-end measure. The peak-end measure specifically models how users remember different moments in their overall experience while listening to a playlist [26]. However, peak-end resulted in similar effects as using a standard mean-aggregation rating.

Regardless of the explanation, the results show that playlist attractiveness is not primarily related to the likeability of its tracks but that other factors such as diversity can play a role.

7 CONCLUSION AND FUTURE WORK

While playlist evaluations can be partly predicted by evaluations of its tracks, other factors of the playlist are more predictive. People seem to evaluate playlists on other aspects than merely its tracks. Even when individual tracks were rated positively, the playlist attractiveness could be low.

We found that both diversity and recommendation approach affected playlist attractiveness. Diversity had a negative effect on playlist attractiveness in recommenders using a low-spread methodology. The track ratings were the most predictive for the playlist attractiveness in the recommendation approach based on genre similarity. Furthermore, inclusion of the highest and last track evaluation score (peak-end) was sufficient to predict playlist attractiveness, performing just as well as the mean of the ratings.

When evaluating recommendation approaches in music recommenders, it is important to consider which evaluation metric to use. Music is often consumed in succession leading to many factors other than track likeability that may influence whether people have satisfactory experiences. Although individual track evaluations are often used in recommender evaluation, track evaluations do not seem to predict playlist attractiveness very consistently.

While we showed that playlist attractiveness is not primarily related to track evaluations, we were unable to effectively measure why certain algorithms generated more attractive playlists compared to others. This question will be addressed in future work. We intent to include a subjective measure for track familiarity. Furthermore, we will identify and attempt to separate distinct musical styles within user preferences. For example, we could give users control about which top artists or top tracks they would like to use to generate recommendations as in [11] to separate the tracks and artists they like under different context.

REFERENCES

[1] Luke Barrington, Reid Oda, and Gert RG Lanckriet. 2009. Smarter than Genius? Human Evaluation of Music Recommender Systems.. In *ISMIR*, Vol. 9. Citeseer, 357–362.

[2] 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.

[3] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. 2010. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 63–70.

[4] Òscar Celma and Perfecto Herrera. 2008. A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 179–186.

[5] Zhiyong Cheng. 2011. Just-for-Me : An Adaptive Personalization System for Location-Aware Social Music Recommendation Categories and Subject Descriptors. (2011).

[6] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User perception of differences in recommender algorithms. *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14* (2014), 161–168. https://doi.org/10.1145/2645710.2645737

[7] Bruce Ferwerda, Mark P Graus, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2017. How item discovery enabled by diversity leads to increased recommendation list attractiveness. In *Proceedings of the Symposium on Applied Computing*. ACM, 1693–1696.

[8] Sophia Hadash. 2019. *Evaluating a framework for sequential group music recommendations: A Modular Framework for Dynamic Fairness and Coherence control.* Master. Eindhoven University of Technology. https://pure.tue.nl/ws/portalfiles/portal/122439578/Master_thesis_shadash_v1.0.1_1_.pdf

[9] Shobu Ikeda, Kenta Oku, and Kyoji Kawagoe. 2018. Music Playlist Recommendation Using Acoustic-Feature Transition Inside the Songs. (2018), 216–219. https://doi.org/10.1145/3151848.3151880

[10] Tristan Jehan and David Desroches. 2004. *Analyzer Documentation [version 3.2].* Technical Report. The Echo Nest Corporation, Somerville, MA. http://docs.echonest.com.s3-website-us-east-1.amazonaws.com/_static/AnalyzeDocumentation.pdf

[11] Yucheng Jin, Bruno Cardoso, and Katrien Verbert. 2017. How do different levels of user control affect cognitive load and acceptance of recommendations?. In *CEUR Workshop Proceedings*, Vol. 1884. CEUR Workshop Proceedings, 35–42.

[12] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. (2018), 13–21. https://doi.org/10.1145/3240323.3240358

[13] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[14] Iman Kamehkhosh and Dietmar Jannach. 2017. User Perception of Next-Track Music Recommendations. (2017), 113–121. https://doi.org/10.1145/3079628.3079668

[15] 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.

[16] Arto Lehtiniemi and Jukka Holm. 2011. Easy Access to Recommendation Playlists: Selecting Music by Exploring Preview Clips in Album Cover Space. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia* (2011), 94–99. https://doi.org/10.1145/2107596.2107607

[17] Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. 2018. Controlling Spotify Recommendations. (2018), 101–109. https://doi.org/10.1145/3209219.3209223

[18] Daniel Müllensiefen, Bruno Gingras, Lauren Stewart, and Jason Ji. 2013. *Goldsmiths Musical Sophistication Index (Gold-MSI) v1.0: Technical Report and Documentation Revision 0.3.* Technical Report. Goldsmiths University of London, London. https://www.gold.ac.uk/music-mind-brain/gold-msi/

[19] F. Pachet, G. Westermann, and D. Laigre. 2001. Musical data mining for electronic music distribution. *Proceedings - 1st International Conference on WEB Delivering of Music, WEDELMUSIC 2001* May 2014 (2001), 101–106. https://doi.org/10.1109/WDM.2001.990164

[20] Steffen Pauws and Berry Eggen. 2003. Realization and user evaluation of an automatic playlist generator. *Journal of new music research* 32, 2 (2003), 179–192.

[21] Alexander Rozin, Paul Rozin, and Emily Goldberg. 2004. The feeling of music past: How listeners remember musical affect. *Music Perception: An Interdisciplinary Journal* 22, 1 (2004), 15–39.

[22] Thomas Schäfer, Doreen Zimmermann, and Peter Sedlmeier. 2014. How we remember the emotional intensity of past musical experiences. *Frontiers in Psychology* 5 (2014), 911.

[23] Markus Schedl, Arthur Flexer, and Julián Urbano. 2013. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems* 41, 3 (2013), 523–539.

[24] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[25] Morgan K Ward, Joseph K Goodman, and Julie R Irwin. 2014. The same old song: The power of familiarity in music choice. *Marketing Letters* 25, 1 (2014), 1–11.

[26] Eelco C. E. J. Wiechert. 2018. *The peak-end effect in musical playlist experiences.* Master. Eindhoven University of Technology.

[27] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modelling and User-Adapted Interaction* 26, 4 (2016), 347–389. https://doi.org/10.1007/s11257-016-9178-6