# User Perception of Next-Track Music Recommendations

Iman Kamehkhosh
TU Dortmund, Germany
iman.kamehkhosh@tu-dortmund.de

Dietmar Jannach
TU Dortmund, Germany
dietmar.jannach@tu-dortmund.de

## ABSTRACT

Many of today's music streaming websites and apps provide personalized next-track listening recommendations based on the user's current and past listening behavior. In the research literature, various algorithmic approaches to determine suitable next tracks can be found. However, almost all of them were evaluated in offline experiments using, for example, manually created playlists as a gold standard. In this work, we aim to check the external validity of insights that are obtained through such offline experiments on historical datasets. We conducted an online user study involving 277 subjects in which the participants evaluated the suitability of four different alternatives of continuing a given set of playlists. Our results indicate that manually created playlists can in fact represent a reasonable gold standard, an insight for which no evidence existed in the literature before. Furthermore, our work was able to confirm that considering playlist homogeneity aspects does not only lead to performance improvements in offline experiments – as indicated by past research – but also to a better quality perception by users. However, the observations also revealed that user studies of this type can be easily distorted by item familiarity biases, because the participants tend to evaluate continuation alternatives better when they know the track or the artist.

## CCS CONCEPTS

**•Information systems →Recommender systems;** *Music retrieval;* **•Human-centered computing →User studies;**

## KEYWORDS

Music Recommendation; Perceived Quality; User Study

## 1 INTRODUCTION

Many modern online music streaming services and mobile apps provide the functionality of generating virtually endless playlists of tracks based on the user's recent listening behavior. The problem of determining suitable playlist continuations has also been explored in the academic literature, and a variety of algorithmic approaches was proposed in the last two decades. To determine such *next-track music recommendations*, these algorithms rely on various types of information, including musical features, meta-data, social tags, or the user's current location and context [4, 5, 17, 31].


*UMAP'17, July 9–12, 2017, Bratislava, Slovakia*

DOI: http://dx.doi.org/10.1145/3079628.3079668

Similar to the general field of recommender systems, the evaluation of playlisting algorithms in academia is mainly accomplished through offline experimental designs [25]. A common approach to benchmark different algorithms is to use playlists created by music enthusiasts as a gold standard. Such "hand-crafted" playlists can be easily obtained from music platforms like last.fm in large quantities. A typical experimental procedure is then to hide individual tracks, e.g., the last track, from a given and assumedly well-designed playlist and let the algorithms predict the hidden tracks [3, 17, 24]. Using such a design, standard information retrieval (IR) measures like Recall or the Mean Reciprocal Rank can be applied. In addition, other quality factors can be assessed with quantitative metrics, e.g., the coherence of the recommended tracks with the given playlist in terms of their musical features [22].

So far, however, limited evidence exists that such computational quality measures are correlated with the actual quality perception of music listeners. One main question in that context is if the hand-crafted playlists are truly representative of the tastes of many users, i.e., if the hidden tracks are considered to be suitable continuations by users other than the creator of the playlist. If this is not the case, being able to predict the hidden track can be of limited value.

In other domains, like movie recommendations, different studies indicate that better performance of algorithms in offline experiments, e.g., in terms of prediction accuracy, does not necessarily translate into a better quality perception by the users or into a positive impact on business metrics. Likewise, it is not always fully clear in which cases users appreciate the recommendation of more diverse or novel items [9, 11, 14, 15].

With this work, we aim to assess to what extent the outcomes of offline experiments in the music domain correlate with the users' quality perception.[1] This question regarding the external validity of findings obtained through offline experiments on historical data has, to our knowledge, not been analyzed in the music recommendation literature to a large extent, even though the limitations of offline experiments are also well-known in the music information retrieval literature, see, e.g., [29].

In the remainder of this paper we first review different insights regarding the performance of different playlists algorithms that were obtained from offline experiments. Next, we describe the design of a user study in which the participants assessed the quality of possible playlist continuations generated by different algorithms. We then discuss the results in detail and in particular check the results for possible familiarity biases, i.e., if the participants considered a suggested continuation as a better match when they already knew the track. The work ends with a discussion of previous works and practical implications of our findings.

[1] For a detailed discussion of the perceived quality of recommendations see [27].

# 2 OFFLINE ANALYSES: COMMON SETUPS AND PAST OBSERVATIONS

## 2.1 Common Offline Experiment Setups

In [4], different possible ways of evaluating automated playlist generation algorithms are reviewed. Since field tests (A/B tests) are rarely possible in academic environments and user studies in the music domain require significant efforts [29], most researchers resort to ex-post analyses and simulation experiments based on existing datasets.

*Datasets.* The datasets used in the literature are often collections of listening logs, e.g., of last.fm users [40], or collections of manually created playlists shared by music enthusiasts, e.g., [30]. Publicly shared playlists have the potential advantage that they are in many cases created with much care by music lovers, e.g., with respect to the included artists, track transitions, etc. [10]. Automatically recorded listening logs, in contrast, can be biased by an existing playlisting algorithm implemented on the site (e.g., an automated radio station). Also, they can contain situations where the user listened to an entire album, which means that the order of the tracks is not necessarily determined by considerations regarding, e.g., track transitions, but by other, not music-related factors. In our user study, we therefore rely on hand-crafted playlists as a gold standard for the evaluation.

*Computational Metrics.* Regarding the evaluation procedure and the computational metrics, one common approach in the recent literature is to split the available playlists into training and test sets, and to let different algorithms predict the last, held-out track of each playlist in the test set [4, 17, 33, 41].

Using this setup, typical information retrieval measures can be applied, including in particular Recall (hit rate), which in our case of only one held-out track is proportional to Precision. In addition, one can not only determine whether a playlister managed to predict the right track, but also if it was at least able to predict the correct artist, genre, or topic. Furthermore, one can measure the homogeneity of the recommended tracks, e.g., in terms of the tempo, and their coherence with the last tracks, e.g., in terms of the genre.

In this work, we will focus on the comparison of such IR measures with the users' quality perceptions. Other computational metrics for assessing playlist continuations were proposed in the literature, e.g., based on the average likelihood of tracks appearing together in playlists [29]. Such approaches are however less popular in the literature and have certain limitations as discussed in [4].

## 2.2 Insights from Offline Experiments

Some of the more recent works that relied on the described "hide-last-track" evaluation approach to benchmark playlisting algorithms include [4, 17, 21, 24] or [22].

In [4], a number of algorithms were compared using different datasets of publicly shared playlists. The set of algorithms included two different rule-mining approaches (association rules and sequential patterns), a k-nearest-neighbor approach, as well as two baseline strategies that recommend the most popular tracks of a selected set of artists. The first of these baselines simply recommends the greatest hits of the artists that appear in the user's most recent listening history. The second, in addition, recommends the most popular tracks of similar artists.

The two following main observations were made:

i) The k-nearest-neighbor (*kNN*) approach led to the best accuracy results across all datasets, when a small $n$ was chosen when determining Recall@n. This situation corresponds to the problem of determining next-track recommendations. Using the nearest neighbors is also favorable in terms of the Recall compared to other, more complex techniques like Bayesian Personalized Ranking (BPR) [36], a comparably recent learning-to-rank method, which is optimized for implicit feedback recommendation scenarios.

ii) A newly proposed simple method called "Collocated Artists – Greatest Hits" (*CAGH*), which recommends the most popular tracks of artists that have often been listened to together in the recent past, however, also led to competitive results. When very long recommendation lists were considered, the method was even consistently better than the *kNN* method.

In other studies [17, 21, 22, 24], the following additional insights were obtained:

iii) Considering additional signals (e.g., musical features, track meta-data, or past listening sessions) in combination with the track co-occurrences captured by the *kNN* method can lead to further accuracy improvements. Incorporating such signals also helps to make the playlist continuations more homogeneous and coherent with the recently listened tracks.

iv) An evaluation using multiple metrics in parallel showed that the comparably simple *CAGH* method mentioned above is particularly competitive in predicting the topic (expressed through tags), genre, or artist of the next track to be played [21].

Given these observations, the goal of our research is to validate the suitability of the offline research setup in general and, more specifically, to test if there is a correspondence between the insights obtained from the offline studies and the subjective quality experience of the users. This leads us to the following research questions, which we aim to answer through a user study.

RQ-1: Are hand-crafted playlists suitable as a gold standard for the evaluation of playlisting algorithms? In other words, will people other than the playlist creator consider the hidden track as a suitable continuation?

RQ-2: Can the consideration of *additional signals*, e.g., musical features or meta-data into the recommendation process improve the perceived quality of the next-track recommendations? Do the observed effects depend on the specific underlying *theme* of a playlist?

RQ-3: How do approaches like *CAGH*, which focus on the most popular items of the artists and which achieve good results in offline experiments, fare in terms of the users' subjective quality perception? Would it be a "safe" strategy to use *CAGH* in practical applications?

RQ-4: Since item familiarity can be a confounding factor in recommender systems user studies that involve popularity-based baselines (like *CAGH*) [23], we ask: Do study participants consider recommendations as more suitable when they already know (and like) the tracks?

# 3 USER STUDY DESIGN

## 3.1 General Procedure for Participants

We created an online application for the purpose of our user study.[2] The participants of the study, which were recruited via different mailing lists, were guided by the application through multiple steps. In an initial step, the application displayed a welcome screen, on which the general purpose of music recommender systems was briefly explained. Then, the participants had to accomplish the following four tasks.

(1) First, the users had to listen to 30-second excerpts of a four-item playlist.[3] The user interface consisted of four vertically aligned audio controls to start and stop the playback of each track. No information about the artists or the tracks themselves was displayed.
(2) When the participants had listened to the tracks, they were asked five questions about the similarity of the songs of the playlist in different dimensions. The questions were related to the emotion, energy, theme, genre, and tempo of the tracks. Using 7-point Likert scale items, the participants could state their degree of agreement (from "not at all" to "completely") that "all songs have the same tempo", etc. The sequence of the questions was randomized across participants. When answering the questions, the participants could listen to the tracks again if they wanted to.[4]
(3) The participants were then presented with four alternative playlist continuations, which were displayed in randomized order across participants. Again, the participants could listen to excerpts of 30 seconds and then state on a 7-point Likert scale item to what extent they agree that the track suits the given playlist (from "not at all" to "perfectly"). In addition, they could indicate if they liked the song (yes, no, indifferent) and if they know the track's artist, the track itself, or both. Again, no information about the recommended tracks was displayed. A screen capture of this part of the application is shown in Figure 1. At all times, the participants could listen to all tracks of the playlist individually or listen to a 30-second summary of all tracks.
(4) In the last step, the participants were asked questions regarding their age group and their music experience. A 7-point Likert scale item ("I randomly listen to music" to "I am a professional") was used for the latter response item. The participants could finally leave some comments and then submit their answers.

The participants could repeat this trial consisting of four steps for up to five different playlists. We discuss the selection of the playlists next.

## 3.2 Selection of Playlists

We used five different playlists in the experiment, which we obtained from three different music platforms (*artofthemix.org*, *last.fm*, *8tracks.com*).

In theory, every participant could have evaluated the possible continuations for all five playlists. In reality, however, most participants completed only one trial and in the end 300 trials were completed by 277 participants. The assignment of the trials was mainly based on a round-robin scheme across the participants to obtain an even number of samples for each playlist. The assignment of the playlist of the first trial was therefore a random one from the participant's perspective.

Since one of our goals was to assess if the choice of the most suitable next track is influenced by certain characteristics of the playlists, we selected the five playlists in a way that each one was very homogeneous in exactly one dimension. We determined the homogeneity of the tracks in the respective dimension by examining the musical features of the tracks as provided by *the.echonest.com*, by considering the social tags assigned by the users to the tracks on last.fm, by analyzing the tracks' lyrics, and by using artist meta-data and expert judgments. We used the following set of playlists.

(1) *Topic-playlist*: This playlist was organized by its creator around the topic *Christmas*. The tracks of the playlist are Christmas pop songs from the 1970s and 1980s.
(2) *Genre-playlist*: This playlist contained tracks of the genre "Soul". We retrieved the genre information of the track artists from *the.echonest.com*.
(3) *Mood-playlist*: The mood playlist contains *romantic* songs. We relied on a semi-automatic analysis of the song lyrics to validate the homogeneity of the playlist in this respect by comparing the TF-IDF (Term Frequency/Inverse Document Frequency) vectors of the tracks of the playlist.
(4) *Tempo-playlist*: The tracks of this playlist all have a similar tempo. The average tempo is 125 beats per minute (bpm) with a standard deviation of only 2 bpm.
(5) *Energy-playlist*: This playlist is homogeneous with respect to the *energy* value provided by the API of *the.echonest.com*. The energy value is based on various types of information, including the loudness of the track. The average energy value of the tracks was at 0.86 on a scale of 0 to 1, with a standard deviation of only 0.02.

When selecting the playlists, we tried our best to make sure that each playlist was homogeneous only in one dimension. We furthermore only selected playlists that contained at most one track per artist. Finally, we commonly avoided choosing playlists that a) contain generally very popular tracks and artists or b) were extreme, e.g., in terms of the tempo. We used the number of occurrences of artists and tracks in the datasets to determine their general popularity level. The detailed descriptions of all playlists can be found online.[5]

Overall, the selection of five playlists with varying characteristics shall help us to test the quality perception for a broader set of *types* of playlists. From the finally selected playlists, we retained only the first five tracks, of which the last one was hidden in the experiment.

## 3.3 Selection of Recommendation Techniques

In each trial, the study participants were presented with four alternative tracks to continue the given playlist. One of the options was the track that was chosen by the playlist creator, i.e., the hidden

[2] The survey can be accessed at https://ls13ap85.cs.tu-dortmund.de:8443/music-survey
[3] The previews were taken from the music service of Spotify. The excerpts were usually not the first 30 seconds of the tracks.
[4] Since these questions refer to the given playlist, we ask them immediately after the participants had listened to the tracks (and before presenting the recommendations) to avoid confusion.
[5] http://ls13-www.cs.tu-dortmund.de/homepage/umap2017-music/playlists.pdf

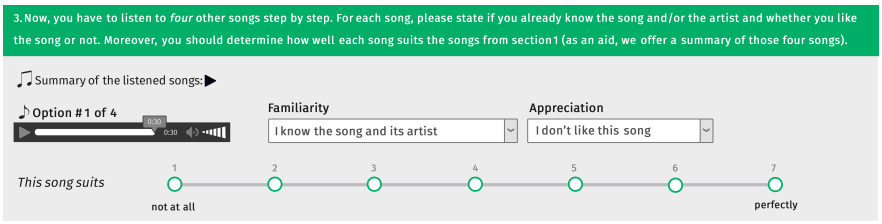


**Figure 1: Evaluating one of four possible continuations.**

track. The other three alternatives were automatically determined using playlisting algorithms from the literature and were presented in randomized order. To train the algorithms we used the dataset from which the playlist was originally taken.[6]

*Selection Rationale.* The following algorithms were included in the experiment. Technical details will be given below.

- *kNN*: A nearest-neighbor method used, e.g., in [4], which performs generally well in terms of the hit rate and which is used as a baseline in a number of research works.
- *CAGH*: A method that selects the greatest hits of certain artists, and which therefore recommends comparably popular items [3, 21]. Including this method helps us also to assess the effects of recommending popular tracks to some extent.
- *kNN+X*: A hybrid method that uses *kNN* as a baseline but considers additional (musical) features to improve the prediction accuracy and to increase the homogeneity of the playlists [24]. The inclusion of this method in the experiment helps us assess the value of considering additional features with respect to the users' quality perception.

*Algorithm Details.* The general task of next-track recommendation techniques is to determine the relevance of a *target track* $t^*$ with respect to a given playlist beginning (listening history) $h$.

*3.3.1 kNN.* The used *kNN* method takes the playlist beginning $h$ as an input and identifies other playlists in the training data that contain the same tracks. The main assumption is that if there are additional tracks in a similar past session (i.e., the "neighbors"), chances are good that these tracks suit the current listening history $h$, too.

Technically, given a listening history $h$, we first compute the binary cosine similarity of $h$ and the other sessions from the training data. The similarity values are then sorted and a set $N_h$ of nearest neighbor sessions of $h$ is determined. The *kNN* score of a target track $t^*$ is then computed as the sum of the similarity values of $h$ and neighbor sessions $nbr \in N_h$ which contain $t^*$ (Equation 1). Note that the indicator function $1_{nbr}(t^*)$ returns 1 if $nbr$ contains $t^*$ and 0 otherwise.

$$score_{\text{kNN}}(h, t^*) = \sum_{nbr \in N_h} sim_{\text{cosine}}(h, nbr) \cdot 1_{\text{nbr}}(t^*) \qquad (1)$$

In our experiments, we set *k=300* to determine the playlist continuations, a parameter that was used for the same datasets also in [24].

[6]The public playlist collections from *last.fm* and *artofthemix.org* that were used in the experiments can be found at http://ls13-www.cs.tu-dortmund.de/homepage/umap2017-music/datasets.zip.

*3.3.2 CAGH.* The *CAGH* method recommends the greatest hits of artists that either appear in the playlist beginning $h$ or are similar to the artists in $h$. The similarity of two artists is estimated based on artist co-occurrences in the playlists. Technically, given a user's current listening history $h$, the *CAGH* algorithm computes the relevance score of a target track $t^*$ by means of Equation 2 .

$$score_{\text{CAGH}}(h, t^*) = \sum_{b \in A_h} sim_{\text{artist}}(a_{t^*}, b) \cdot cnt(t^*) \qquad (2)$$

$A_h$ is the set of artists in current history, $cnt(t^*)$ is the number of occurrences of $t^*$ in the training data, which corresponds to the most frequently occurring tracks of the respective artists in the dataset, and $a_{t^*}$ is the artist of the target track. As a measure of similarity of two artists $sim_{\text{artist}}(a_{t^*}, b)$, we count how often two artists appear together in the sessions of the training set.

*3.3.3 Hybrid: kNN+X.* In our experiment, we use the scoring scheme proposed in [24] to combine the scores returned by the *kNN* method with a score that expresses the suitability of a candidate track in terms of a certain (musical) feature.

Given a scoring function of a feature $score_{\text{f}}$, the combined score is computed as

$$score_{\text{hybrid}}(h, t^*) = \alpha \cdot score_{\text{kNN}}(h, t^*) + (1 - \alpha) \cdot score_{\text{f}}(h, t^*) \quad (3)$$

where the weight factor $\alpha$ is used to balance the relative importance of the baseline and the feature scores. The following feature scoring functions were used:

- *Topic-Scorer:* We use the social tags assigned to the tracks to determine the topic of a playlist. As done in [24], we compute TF-IDF vectors for each track using the tags. The average cosine similarity of the TF-IDF vectors of the tracks of a playlist represents the homogeneity level in terms of the topic.

  The topic score $score_{\text{topic}}$ of a target track $t^*$ with the TF-IDF vector $\vec{t^*}$, given a history $h$ consisting of tracks $t_1, \ldots, t_n$ with the TF-IDF vectors $\vec{t_1}$ to $\vec{t_n}$ is computed as follows.

$$score_{\text{topic}}(h, t^*) = sim_{\text{cosine}}\left(\frac{\Sigma_{t_i \in h} \vec{t_i}}{|h|}, \vec{t^*}\right) \qquad (4)$$

- *Genre-Scorer:* To determine the genre of a playlist, we use the genres of the artists of its tracks.[7] The same content-based approach as the one used above for topics can be used for the genres. The TF-IDF vectors for each track are computed using the genres of its artist. The average cosine

[7]Genre information about individual tracks was in most cases not available to us.

similarity of the TF-IDF vectors of the tracks of a playlist is then used to measure its homogeneity level in terms of the genre. Similar to the topic score, Equation 4 can be used for computing the genre score $score_{\text{genre}}$.

- *Mood-Scorer:* Similar to [18], we applied a mood classification method to determine the general mood of the tracks based on their lyrics. First, list of mood-related social tags was used to create an initial ground truth set for a selected number of moods. The lyrics of the ground truth tracks were then processed and TF-IDF vectors were generated. A Support Vector Machines (SVM) classifier was finally applied on the generated TF-IDF vectors to create a binary predictive model for each target mood. These models were then used to predict the mood of unlabeled tracks.

  As a result, each track is labeled with a set $M_t$ of moods. Accordingly, we build a set $M_h$ of moods for each playlist. Each mood in this set is weighted by the number of occurrences of that mood divided by the number of the tracks of the playlist ($w_m$). For a target track $t^*$, the mood score $score_{\text{mood}}$ is then computed as the sum of the weights of the common moods of the target track $M_{t^*}$ and the given history $M_h$, i.e., $1_{M_h}(m) = 1$ if $M_h$ contains the mood of the target track $m$ and 0 otherwise.

$$score_{\text{mood}}(h, t^*) = \sum_{m \in M_{t^*}} w_m \cdot 1_{M_h}(m) \tag{5}$$

- *Tempo and Energy-Scorer:* Musical features such as the tempo or the energy can be relevant factors when selecting tracks, e.g., for a workout or a party playlist. We correspondingly use the proposed scoring scheme in [24] for such numerical features based on the mean $\mu_h$ and standard deviation $\sigma_h$ of the observed values in the given playlist (listening history). Given a history $h$ and a feature value of a target track $f_{t^*}$, the value of the probability density function of a Gaussian distribution is used as a numerical feature score for tempo and energy.

$$score_{\text{numfeature}}(h, t^*) = \frac{1}{\sigma_h \sqrt{2\pi}} e^{-\frac{(f_{t^*} - \mu_h)^2}{2\sigma_h^2}} \tag{6}$$

To compute the scores for our user study, we determined the weight ($\alpha$) for the hybrid scorer by optimizing the hit rate on the training data as done in [24].[8]

# 4 RESULTS AND OBSERVATIONS

## 4.1 Participation Statistics

We recruited study participants by inviting students of university classes in Germany and Brazil, by sending emails to friends and colleagues, and by posting invitations on social network sites. 300 trials were completed by 277 subjects during a period of 8 weeks.[9] Each of the 5 playlists was evaluated 60 times. Most (83%) of the participants were aged between 20 and 40 and the majority of the participants was from Germany (43%) or Brazil (40%). On a scale between 1 and 7, the median of the self-reported experience with music was 5, i.e., the participants were on average quite experienced or interested in music.

The average time for participants to complete one trial was at about 10 minutes. On average, they listened to about 22 seconds (of 30) of each track of the playlist and about 26 seconds of the recommended track. This indicates that the participants completed the survey carefully.

## 4.2 Study Outcomes

In the following, we will discuss the outcomes of the study with respect to the research questions from Section 2.2.

*Determining a Ranking.* The ultimate goal of many existing research papers is to determine a ranking of different algorithms, e.g., with respect to their accuracy. Since our goal is to determine the correspondence of offline results with the results obtained through the user study, we use a ranking-based approach to investigate the different research questions.

We use two different rank-aggregation strategies.[10]

(1) We report how often each recommendation technique was the *winner*, i.e., how often its recommendation was considered the most suitable alternative continuation. In each trial, an alternative was counted as the most suitable, if (a) it was rated higher than the other alternatives and (b) the *suitability* value assigned to it was greater than four. We set this threshold to avoid counting it as a "win" when a recommendation was actually not good, but merely better than other, even worse alternatives. In 10% of the cases, no clear winner could be determined based on these rules.
(2) We apply the Borda Count (BC) rank aggregation strategy to determine the *ranking of all four alternatives*. The responses provided by the participants are used as implicit ranking information. In the Borda Count computation scheme, the highest ranked alternative gets *n-1* points, where *n* is the number of alternatives to be ranked. Each subsequent alternative gets one point less, so that the lowest-ranked alternative gets no points.

The result scores of the four different track continuation alternatives (i.e., the three ranking algorithms and the hidden track) are given in Table 1. To investigate to what extent familiarity aspects may affect the results, we report the results for two different configurations. In the "All Tracks" configuration, we consider the rankings of all trials (which fulfill the above-mentioned criteria). In the "Novel Tracks" configuration, we consider only trials where the participants indicated that they did not already know the most suitable track or its artist, which was the case in about 70% of all trials with a unique winner. To compute the Borda Count for novel tracks, we used a variant of the Borda Count from [12], which is designed for the aggregation of partial rankings ($\text{BC}_{avg}$).

*4.2.1 RQ-1: Are hand-crafted playlists suitable for the evaluation of playlisting algorithms?* In 41% of all trials with a unique winner, the participants selected the track as the most suitable continuation

[8] In the combinations of the *kNN* method with the topic and tempo scorers, $\alpha$ was set to 0.3 and in other combinations it was set to 0.7.
[9] We removed 9 additional participants from the study who did not listen to any track of the given playlist.
[10] The data collected in our study is ordinal, i.e., a ranking of the response levels is possible. However, we cannot assume equidistance between the response levels, which ranged from "(1) not suitable at all" to "(7) perfectly suitable". Using descriptive statistics like the mean and the standard deviation to aggregate the results would therefore be questionable from a methodological perspective.

**Table 1: Ranking results.**

| | All Tracks | | Novel Tracks | |
|---|---|---|---|---|
| Strategy | Ranked #1 | BC | Ranked #1 | $BC_{avg}$ |
| Hidden Track | 41% | 645 | 43% | 580 |
| CAGH | 46% | 649 | 32% | 403 |
| kNN | 25% | 520 | 19% | 477 |
| kNN+X | 30% | 594 | 36% | 631 |

that was originally picked by the creator of the given playlist. The difference between the score of the hidden track and the alternative computed by the *CAGH* method, which focuses on popular tracks that are often already known to the study participants, is not statistically significant.[11] The same ranking of the algorithms is obtained when the Borda Count is used.

When considering only trials in the evaluation where the participants did not know the track or its artist, the hidden track was significantly more often chosen as the most suitable option than the other alternatives; see the right-hand part of Table 1. The hidden track was also a good choice when using the Borda Count method; in this case, however, the *kNN+X* method significantly outperformed the other alternatives.

Overall, we see this as clear indicator that the manually created playlists used in the experiments are of good quality and that it can, in general, be meaningful to use such playlists as gold standard in offline experiments. Assuming that (many) publicly shared playlists are in fact created with care also means that these playlists can be used as a reliable source for detecting hidden patterns and relationships between tracks that can be exploited by next-track recommendation algorithms.

*4.2.2 RQ-2: Can considering additional (musical) signals in the recommendation process improve the users' quality perception?* To answer this question, we can compare the suitability assessments of the methods *kNN* and *kNN+X*. The latter method considers different additional signals in the recommendation process and outperforms the pure *kNN* method in offline experiments. Our results clearly show that considering different additional signals also leads to a statistically significant improvement in the quality (suitability) perception by the study participants. The *kNN+X* method was considered as more suitable than the *kNN* method in both configurations and on both measures, which indicates that the users in fact preferred track continuations that are coherent with the recently played tracks in different dimensions. When considering only novel track recommendations, the *kNN+X* method was actually the best performing one in terms of the Borda Count and the second best one in terms of the other measure (frequency of winning).

Coherence aspects were, however, not equally important for all tested signals. Additional analyses, which we do not report here for space reasons, showed that coherence in terms of the genre and the topic of the playlist was particularly important for the participants, i.e., in these cases, the *kNN+X* recommendation were generally ranked higher than those of the *kNN* method. An exception from this general pattern was observed for the *tempo*-oriented playlist, where the *kNN* and the *CAGH* method were on average ranked higher than the *kNN+X* method. This was unexpected for two reasons. First, the tracks that were recommended by the *kNN* and the *CAGH* method was at about 75 bpm and far away from the average tempo of the given playlist, which was at 125 bpm. Second, 68% of the participants had in fact correctly identified the tempo as an underlying theme of the playlist according to the questionnaire (see Table 2). As a result, this indicates that in some cases homogeneity in terms of the tempo might be less important for users than other characteristics like artist homogeneity. The sometimes limited importance of the tempo was also observed in [20] based on an analysis of a larger pool of public playlists.[12]

*4.2.3 RQ-3: How do popularity-based approaches like CAGH fare in terms of the subjective quality perception by users?* The *CAGH* method not only leads to competitive results in offline experiments [21], its recommendations are also often considered to be very suitable by the participants of our study. This also indicates that using the hit rate *can be* one suitable proxy measure to evaluate different algorithms in offline experiments. However, when we only consider the participants' suitability assessments for tracks that are novel to them, the *CAGH* method does not perform as well as other approaches and even has the lowest performance across all possible alternatives with respect to the Borda Count.

Overall, according to our experiment, the popularity-biased *CAGH* can be considered as a "safe" strategy to determine generally suitable tracks in practice, particularly when it comes to cold start users. The capability of such a technique to help users discover new tracks or artists over time is, however, limited. In cases where item discovery is considered as a key value-adding feature, e.g., of an online streaming service, *CAGH* might therefore not be the best choice.

*4.2.4 RQ-4: Do subjects consider recommendations as more suitable when they already know (and like) the tracks?* With this research question, we aim to understand if the results of studies like ours can be biased by item familiarity effects, i.e., if the study participants tend to like a recommendation when they already know the track. Such an effect was observed in a user study in the movie domain in [23], where the simple recommendation of generally popular movies was the "winning" strategy.

The analyses in the previous sections indeed suggest that such effects can also exist in the music domain. There are measurable differences between the ranking of the alternatives when it comes to known or novel track recommendations. Specifically, the popularity-based *CAGH* method works particularly well when all trials are considered, but users found its recommendations often not appropriate when they did not know the track already. This means that the strong performance of the *CAGH* in our experiment in one configuration might be a light overestimation of the true performance of the method in practical settings.[13]

We analyze the differences between the algorithms in terms of the familiarity of users with their track recommendations in more depth in the following section.

[11] To test for statistical significance, we use the Mann-Whitney U test with $p < 0.05$.

[12] The relevance of different musical aspects can depend on the listening context. The inclusion of tracks in a playlist that all have a similar tempo might for example be desirable when listening to music while doing sports.

[13] The findings reported in [23] are based on a study on Mechanical Turk, and the motivation of the participants to provide reliable answers might be lower than in our experiment with volunteers that had no monetary incentive.

### 4.3 Analysis of Track Familiarity and Affinity

During the user study, when the participants assessed the suitability of a track as a possible continuation for the given playlist, they were asked to indicate if they already knew the recommended track or at least its artist. In addition, they could specify to what extent they *liked* the track, independent of its suitability for the given playlist. The results regarding the users' familiarity with the recommended alternatives are shown in Figure 2.

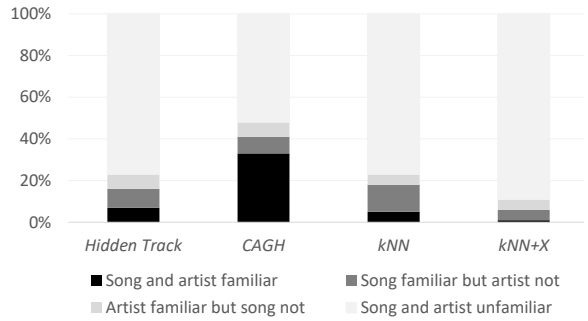


**Figure 2: Familiarity level of alternative next-track recommendations in all trials.**

The *CAGH* method by design recommends popular tracks, and it is not surprising that in almost half of the trials the subjects were already familiar with the tracks in some form. The hybrid *kNN+X* method, which focuses on coherence aspects, in contrast, recommended the largest amount of novel items. In about 90% of the trials, the recommended items were completely unknown to the study participants. However, this tendency of recommending novel tracks did not hurt the quality perception of users when compared with the plain *kNN* method (as shown in Table 1), maybe because the novel track recommendations were coherent with the recent listening history. A different effect was previously observed for the movie domain in [11], where novel item recommendations led to a lower quality perception.

Looking at the like/dislike statistics for the tracks themselves – we omit the detailed results for space reasons – we observed that the tracks that were recommended by the *CAGH* method were liked in about 60% of the cases. For all other alternatives, the percentage was much lower and at about 40% with no significant differences between the *kNN* and the *kNN+X* method. This indicates that the study participants distinguished if they generally liked a certain track or if they considered a track to be a suitable continuation for the given playlist. This is another indication that the study participants answered the questionnaire with care.

### 4.4 Perception of Playlist Characteristics

The results in Section 4.2 showed that the participants found playlist continuations more often suitable when they were coherent with the recently played tracks. The final question we analyze in this paper is to what extent the study participants actually *noticed* that there is an underlying design rationale for the different playlists and if they could identify the underlying theme.

During the experiment, we asked the users to answer to what extent they agree that the tracks of the given playlist had similar characteristics, e.g., in terms of the tempo or mood, see Section 3. The main result of this analysis is shown in Table 2. The first row of the table shows the true underlying design rationale of the playlists. The second row shows the "winner" in terms of the users' perception. Looking, for example, at the column labeled with "Topic", we see that in 95% of the trials, the tracks of the playlist were indeed considered to be more similar to each other in terms of the topic than in any other dimension. Across all trials, the participants correctly identified the true theme of the playlist as a winner in almost two third of the cases.

**Table 2: Frequencies of dominating characteristics as perceived by the participants.**

| True Theme | **Topic** | **Genre** | **Mood** | **Tempo** | **Energy** |
|---|---|---|---|---|---|
| Subjective Perception | Topic (95%) | Genre (55%) | Topic (40%) | Tempo (68%) | Genre (48%) |

In two cases, the participants, however, considered other themes as being more important than the true one. In the case of the *mood*-playlist, the participants more often found the tracks to be more similar in terms of the topic (40%) than in terms of the mood. About 30% of the users (not reported in the table) found the mood to be the dominating feature. Since the mood coherence of the playlist was identified by us based on the lyrics of the tracks, we believe that the fact that the participants could only listen to 30-second excerpts and probably did not focus on the lyrics contributed to this result. The *energy*-based playlist was the second one where we observed deviations. 48% of the participants rather felt that the tracks are connected by the genre than by the energy level (35%, not reported in the table). The reasons could be that some of the participants were less familiar with the term energy than with other concepts or that the energy level, as computed by *the.echonest.com*, does not always correlate well with the users' perception.

## 5 RESEARCH LIMITATIONS

Evaluating recommenders with user studies is challenging in different ways. In the music domain, the participants have to invest a considerable amount of time as they are required to listen to a number of tracks during the experiment. In our setup, we therefore limited the number of tracks in the given playlist to four and only provided 30-second previews of the tracks. Since we used excerpts that were selected and provided by a commercial service (Spotify), we are confident that the excerpts are representative of the tracks. Furthermore, all of the tracks in the experiment were not longer than usual pop songs and each track exhibited limited within-track variation of the tempo or harmonics.

Another challenge is that recommending mostly popular items might lead to a familiarity bias, i.e., the participants tend to rate items they already know highly [23] and dislike items they do not know, as was observed in [11] for the movie domain. In our experiment, we therefore selected playlists that did not contain too popular tracks and we did not reveal additional track or artist information. Furthermore, we explicitly asked the participants to indicate whether or not they knew the track already to quantify possible biases.

A more general limitation of laboratory studies is that when users feel being supervised or in a "simulation" mode, they might behave differently than when they are within one of their normal music

listening environments. To alleviate this problem, we provided an online application to enable users to participate in the study when and where they wanted to.

Academic user studies in the music domain often have a limited size and in many cases only involve 10 to 20 participants in total [4]. Our study involved 277 participants and 300 trials, leading to 60 trials per condition. The majority of the study participants were university students in Germany and Brazil. While this population of digital natives might be representative of many users of today's digital music services, it still has to be shown to what extent the findings of the study generalize to other types of music listeners and to other types of musical genres.

Finally, our study was based on a specific collection of five tracks and we have to be aware that the choice of the playlists might have impacted the observed outcomes. To minimize this threat to validity, we have selected the playlists used in the experiments in a way that they (a) cover a broader range of music preferences, and (b) that they were assembled by their creators with different design rationales in mind. We, however, limited the number of playlists in the experiment to five in order to end up with a sufficient number of participants per playlist.

## 6 RELATED WORKS

Laboratory or online studies on music recommendation, and in particular on playlist generation, are comparably rare. The work by Barrington et al. [2] is one of the few exceptions. Similar to our work, they compare different music playlisting approaches in a user study. Their set of compared algorithms includes Apple's Genius collaborative filtering based system and, among others, a method based on artist similarity. As part of their experiment, they revealed the artist and track information in one condition and did not disclose this information in another. An interesting insight was that the Genius system was "overwhelmingly superior" when no information was displayed, whereas the artist-based method was perceived to produce slightly better recommendations in the other case. This effect is to some extent related to our observation regarding possible familiarity effects that can influence the users' evaluation of a recommendation. Overall, however, the main goal of the work by Barrington et al. was to compare content-based and collaborative filtering based methods. Our main goal, in contrast, was to validate the results obtained in offline experiments as was done for other recommendation domains in [8, 14, 19, 26] or [37].

A smaller number of user studies can be found in the literature that focus on specific aspects of different music recommendation scenarios. There are works that investigate, e.g., the role of track orders, track positions, and the problem of recommending *collections of items* [16]; works on the recommendation or selection of a suitable playlist for *groups of users* [34]; works on location-based contextualized music recommendation [5]; works that analyze the effect of visualizations and user control on the user experience [1]; and works that address the question of how different personality traits of users impact their perception of music [6, 13, 32]. In our study, we in contrast focused on a very general recommendation scenario and the comparison of offline algorithm performance and the subjective user experience.

Finding new performance measures that correlate with the subjective user experience was the goal of the recent work by Craw et al. [7]. In their work, a new computational metric is proposed that balances listening events and explicit ratings to avoid that the recommendation of less popular items is "punished". They validated that their metric corresponds to the perceived quality assessment by music listeners through a user study involving 132 subjects. We see their work as complementary to ours, as we were not interested in defining new measures but in the assessment of the usefulness of commonly used IR measures.

Such IR measures, while often used in the literature, are not the only computational metrics that can be used to assess the quality of playlists. One can, for example, compute the diversity of a playlist (e.g., in terms of the genres), determine the coherence of the individual tracks in various dimensions, or analyze the smoothness of the transitions between two consecutive tracks. The main question when relying on such computational metrics, however, is if they are representative of the users' quality perception. One way to investigate this aspect is to analyze hand-crafted playlists in order to see if music enthusiasts follow certain (implicit) guidelines when they create playlists and, for example, select tracks for inclusion that are homogeneous in certain respects or not. Examples of works that investigated such patterns in playlists are [10, 20, 38] and [39]. The existence of such patterns indicates that certain quality characteristics, e.g., diversity, *can* in general be relevant when making playlist recommendations, but users might have different preferences regarding how diverse a playlist should be [39]. Furthermore, the relevance of certain aspects can depend on the contextual situation of the user. In our work, we therefore focused on traditional IR measures and consider the investigation of these other aspects as important areas for future work.

Overall, we see our work as one further contribution in the context of user-centered evaluation approaches for recommendation systems, which have gained increasing popularity in recent years and which led to the development of new evaluation frameworks [27, 35]. For the domain of Music Information Retrieval, Lee and Price [28] recently discussed the limitations of the current research practice in the field and stressed the importance of user-centered evaluation approaches. They also conducted a qualitative user study, where the goal was to better understand the various factors that can have an influence on the users' quality perception of (commercial) music services. The overall goal of their research is to develop a more comprehensive evaluation approach that considers a variety of relevant factors beside accuracy, including the user interface design or privacy and trust aspects.

## 7 CONCLUSIONS

In this work, we have investigated the quality perception of playlist continuation proposals generated by different next-track music recommendation techniques. Since several observations obtained in offline experiments could be reproduced in the user study, one main insight of the work is that hand-crafted playlists shared by music enthusiasts can indeed be a valuable basis for designing and evaluating recommendation algorithms. Furthermore, our work provided evidence that recent approaches that focus both on playlist coherence and prediction accuracy not only lead to better results in offline experiments, but also lead to an improved quality perception by users.

## REFERENCES


[1] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive Mood-based Music Discovery and Recommendation. In *UMAP '16*. 275–279.

[2] Luke Barrington, Reid Oda, and Gert R. G. Lanckriet. 2009. Smarter than Genius? Human Evaluation of Music Recommender Systems. In *ISMIR '09*. 357–362.

[3] Geoffray Bonnin and Dietmar Jannach. 2013. Evaluating the Quality of Playlists Based on Hand-Crafted Samples. In *ISMIR '13*. 263–268.

[4] Geoffray Bonnin and Dietmar Jannach. 2014. Automated Generation of Music Playlists: Survey and Experiments. *ACM Computing Surveys* 47, 2 (2014), 26:1–26:35.

[5] Matthias Braunhofer, Marius Kaminskas, and Francesco Ricci. 2013. Location-aware Music Recommendation. *International Journal of Multimedia Information Retrieval* 2, 1 (2013), 31–44.

[6] Pei-I Chen, Jen-Yu Liu, , and Yi-Hsuan Yang. 2015. Personal Factors in Music Preference and Similarity: User Study on the Role of Personality Traits. In *CMMR '15*.

[7] Susan Craw, Ben Horsburgh, and Stewart Massie. Music Recommenders: User Evaluation Without Real Users?. In *IJCAI '15*. 1749–1755.

[8] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Papadopoulos, and Roberto Turrin. 2011. Comparative Evaluation of Recommender System Quality. In *CHI EA'11*. 1927–1932.

[9] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2012. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study. *Transactions on Interactive Intelligent Systems* 2, 2 (2012), 11:1–11:41.

[10] Sally Jo. Cunningham, David. Bainbridge, and Annette. Falconer. 2006. 'More of an Art than a Science': Supporting the Creation of Playlists and Mixes. In *ISMIR '06*. 240–245.

[11] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *RecSys '14*. 161–168.

[12] Peter Emerson. 2013. The Original Borda Count and Partial Voting. *Social Choice and Welfare* 40, 2 (2013), 353–358.

[13] Bruce Ferwerda, Mark Graus, Andreu Vall, Marko Tkalčič, and Markus Schedl. 2016. The Influence of Users' Personality Traits on Satisfaction and Attractiveness of Diversified Recommendation Lists. In *EMPIRE '16 Workshop at RecSys '16*. 43–47.

[14] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and Online Evaluation of News Recommender Systems at swissinfo.ch. In *RecSys '14*. 169–176.

[15] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems* 6, 4 (2015), 13:1–13:19.

[16] Derek L. Hansen and Jennifer Golbeck. 2009. Mixing It Up: Recommending Collections of Items. In *CHI '09*. 1217–1226.

[17] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2012. Context-Aware Music Recommendation Based on Latent Topic Sequential Patterns. In *RecSys '12*. 131–138.

[18] Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann. 2009. Lyric Text Mining in Music Mood Classification. In *ISMIR '09*. 411–416.

[19] Dietmar Jannach and Kolja Hegelich. 2009. A Case Study on the Effectiveness of Recommendations in the Mobile Internet. In *RecSys '09*. 205–208.

[20] Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. 2014. Analyzing the Characteristics of Shared Playlists for Music Recommendation. In *RSWeb '14 Workhop at RecSys '14*.

[21] Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. 2016. Biases in Automated Music Playlist Generation: A Comparison of Next-Track Recommending Techniques. In *UMAP '16*. 281–285.

[22] Dietmar Jannach, Iman Kamehkhosh, and Lukas Lerche. 2017. Leveraging Multi-Dimensional User Models for Personalized Next-Track Music Recommendation. In *SAC '17*.

[23] Dietmar Jannach, Lukas Lerche, and Michael Jugovac. 2015. Item familiarity as a possible confounding factor in user-centric recommender systems evaluation. *i-com Journal for Interactive Media* 14, 1 (2015), 29–40.

[24] Dietmar Jannach, Lukas Lerche, and Iman Kamehkhosh. 2015. Beyond "Hitting the Hits": Generating Coherent Music Playlist Continuations with the Right Tracks. In *RecSys '15*. 187–194.

[25] Dietmar Jannach, Markus Zanker, Mouzhi Ge, and Marian Gröning. 2012. Recommender Systems in Computer Science and Information Systems–A Landscape of Research. In *EC-Web '12*. 76–87.

[26] Evan Kirshenbaum, George Forman, and Michael Dugan. 2012. A Live Comparison of Methods for Personalized Article Recommendation at Forbes.com. In *ECML/PKDD '12*. 51–66.

[27] 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* 4-5 (2012), 441–504.

[28] Jin Ha Lee and Rachel Price. 2016. User experience with commercial music services: An empirical exploration. *Journal of the Association for Information Science and Technology* 67, 4 (2016), 800–811.

[29] Brian McFee and Gert R. G. Lanckriet. 2011. The Natural Language of Playlists. In *ISMIR '11*. 537–542.

[30] Brian McFee and Gert R. G. Lanckriet. 2012. Hypergraph Models of Playlist Dialects. In *ISMIR '12*. 343–348.

[31] Joshua L Moore, Shuo Chen, Thorsten Joachims, and Douglas Turnbull. 2012. Learning to Embed Songs and Tags for Playlist Prediction. In *ISMIR '12*. 349–354.

[32] Melissa Onori, Alessandro Micarelli, and Giuseppe Sansonetti. 2016. A Comparative Analysis of Personality-Based Music Recommender Systems. In *EMPIRE '16 Workshop at RecSys '16*. 55–59.

[33] John C. Platt, Christopher J. C. Burges, Steven Swenson, Christopher Weare, and Alice Zheng. 2001. Learning a Gaussian Process Prior for Automatically Generating Music Playlists. In *NIPS '11*. 1425–1432.

[34] George Popescu and Pearl Pu. 2012. What's the Best Music You Have?: Designing Music Recommendation for Group Enjoyment in Groupfun. In *CHI EA '12*. 1673–1678.

[35] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *RecSys '11*. 157–164.

[36] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI '09*. 452–461.

[37] Marco Rossetti, Fabio Stella, and Markus Zanker. 2016. Contrasting Offline and Online Results when Evaluating Recommendation Algorithms. In *RecSys '16*. 31–34.

[38] Andy M. Sarroff and Michael Casey. 2012. Modeling and Predicting Song Adjacencies In Commercial Albums. In *SMC '12*.

[39] Malcolm Slaney and William White. 2006. Measuring Playlist Diversity for Recommendation Systems. In *AMCMM '06*. 77–82.

[40] Roberto Turrin, Massimo Quadrana, Andrea Condorelli, Roberto Pagano, and Paolo Cremonesi. 2015. 30Music Listening and Playlists Dataset. In *Poster Proceedings RecSys '15*.

[41] Linxing Xiao, Lie Lu, Frank Seide, and Jie Zhou. 2009. Learning a Music Similarity Measure on Automatic Annotations with Application to Playlist Generation. In *ICASSP '09*. 1885–1888.