UMUAI manuscript No.
(will be inserted by the editor)

Effects of Recommendations on the Playlist Creation Behavior of Users

Iman Kamehkhosh · Geoffray Bonnin · Dietmar Jannach

Received: date / Accepted: date

Abstract The digitization of music, the emergence of online streaming platforms and mobile apps have dramatically changed the ways we consume music. Today, much of the music that we listen to is organized in some form of a playlist, and many users of modern music platforms create playlists for themselves or to share them with others. The manual creation of such playlists can however be demanding, in particular due to the huge amount of possible tracks that are available online. To help users in this task, music platforms like Spotify provide users with interactive tools for playlist creation. These tools usually recommend additional songs to include given a playlist title or some initial tracks. Interestingly, little is known so far about the effects of providing such a recommendation functionality. We therefore conducted a user study involving 270 subjects, where one half of the participants – the treatment group – were provided with automated recommendations when performing a playlist construction task. We then analyzed to what extent such recommendations are adopted by users and how they influence their choices. Our results, among other aspects, show that about two thirds of the treatment group made active use of the recommendations. Further analyses provide additional insights about the underlying reasons why users selected certain recommendations. Finally, our study also reveals that the mere presence of the recommendations impacts the choices of the participants, even in cases when none of the recommendations was actually chosen.

Iman Kamehkhosh
E.ON, Germany
E-mail: iman.kamehkhosh@eon.com

Geoffray Bonnin
Loria, Nancy, France
E-mail: geoffray.bonnin@loria.fr

Dietmar Jannach
University of Klagenfurt, Austria
E-mail: dietmar.jannach@aau.at

1 Introduction

In recent years, music streaming has become the predominant way of consuming music and the most profitable source in the music industry (Friedlander, 2017). At the same time, the consumption of music on modern online streaming sites has become largely determined by *playlists*, i.e., sequences of tracks that are intended to be listened to together (Bonnin and Jannach, 2014; Schedl et al., 2018). In fact, according to Nielsen's 2017 Music survey (Nielsen, 2017), almost three fourth (74 %) of the music experience of the users is based on playlists. Furthermore, more than half of the survey respondents stated that they create playlists by themselves.

Creating playlists can, however, be a laborious task. In principle, many factors can potentially be considered in the creation process like the transitions between the tracks or the musical or thematic coherence of the selected musical pieces (Cunningham et al., 2006; Hagen, 2015). The availability of millions of tracks on today's music platforms like Spotify adds further complexity to the task. In general, a large set of options can soon become overwhelming for the users, and finally lead to problems of "overchoice" or "choice overload" (Iyengar and Lepper, 2000; Gourville and Soman, 2005; Scheibehenne et al., 2010). As a result, users can find it more and more difficult to choose any of the options. In the case of playlist creation, although users only know a subset of the available options, they still perceive the identification of suitable tracks from memory as being difficult, and often wish to extend their music selection by discovering new tracks (Hagen, 2015).

One possible way of assisting users in the playlist creation process is to provide them with automatically generated track suggestions through a recommendation service. How to automatically select appropriate items for recommendation given an initial set of tracks has been extensively studied in the literature (Bonnin and Jannach, 2014). Furthermore, playlist creation support tools can nowadays also be found in practice. Spotify, for example, as one of the market leaders in the area of online music services, provides such a functionality to its users.

As in other application areas of recommender systems, academic research in the field is to a large extent based on offline experimental designs and focused on optimizing the prediction accuracy of machine learning models (Jannach et al., 2012). To what extent higher prediction accuracy translates into higher utility for the user, to higher user satisfaction, or to increased adoption of the service is, however, not always fully clear (Jones, 2010). In fact, many factors can be relevant for the adoption of a recommendation service, including the ease of use of the interface, the users' trust in the system as well as recommendation-related aspects like the diversity of the track suggestions (Jones, 2010; Armentano et al., 2015; Tintarev et al., 2017).

Given this research gap, the goal of the work presented in this paper is to go beyond algorithmic approaches for track selection and to explore research questions related to the adoption and influence of recommender systems during the playlist creation task. To that purpose, we have conducted a controlled

user study involving 270 subjects, where the task of the participants was to create a playlist for one of several predefined topics. About half of the participants – the treatment group – were supported by a recommender system, which used two different strategies for track selection. The main goals of our study were to understand (i) to what extent and for what reasons system-generated recommendations are actually adopted by users, and (ii) how the recommendations influenced the playlist creation behavior of the users.

One main finding of our work is that track recommendations were highly adopted by the participants. More than 67 % of the participants of the treatment group picked at least one recommended track for inclusion in their playlists, which we see as a strong indicator of the utility of the provided functionality. The recommendations furthermore had a positive effect on some other dimensions. In particular, they helped participants discover relevant tracks and influenced their choices even when they did not select one of the recommendations. Overall, our study provides a number of novel insights on the effects of recommendations on users in the music domain. It emphasizes the practical relevance of such systems and leads to a number of implications regarding the design of such systems. Generally, we therefore see our work as an important step towards more comprehensive and user-centric approaches for the evaluation of music recommenders (Pu et al., 2011; Knijnenburg and Willemsen, 2015).

The paper is organized as follows. Section 2 provides a brief review of related works and previous user studies on the topic. Further in the same section, we summarize the insights of an exploratory study as described in (Kamehkhosh et al., 2018). We present our research questions in Section 3 and provide the details of the main study in Section 4. The outcomes of the study are discussed in Section 5. Finally, in Section 6, we discuss the practical implications that result from our observations, list the research limitations and threats to validity of the study and present an outlook on future works.

2 Previous Works

2.1 Algorithms for Next-Track Music Recommendation

The majority of research works in the area of recommender systems, as mentioned above, aims to improve the prediction accuracy of an algorithm based on long-term user profiles. In the specific problem setting in our scenario – the recommendation of additional tracks during playlist construction – we, however, do not make the assumption that such a user profile is available for personalization.

Differently from the traditional "matrix completion" problem formulation, the computational task in our case is thus to determine a set of suitable musical tracks given a sequence of previous tracks. A variety of algorithms have been proposed for this scenario that is commonly referred to as "next-track music recommendation", "playlist continuation" or "session-based recommendation"

(Bonnin and Jannach, 2014; Quadrana et al., 2018). The proposals range from the application of sequential pattern mining, session-based nearest neighbor techniques, random walks on a hypergraph whose edges correspond to playlist categories, hybrid techniques based on meta-data and content information, re-ranking techniques based on musical features, or neural networks that leverage various types of data (Hariri et al., 2012; McFee and Lanckriet, 2012; Jannach et al., 2015; Vall et al., 2018).

A comparative evaluation in (Bonnin and Jannach, 2014) revealed that in terms of typical information retrieval (IR) measures like the recall, a newly proposed method called CAGH was particularly competitive. The conceptually simple method ("Collocated Artists – Greatest Hits") takes the artists of the given playlist beginning as an input and recommends the most popular tracks of these artists and of other artists that are similar to those appearing in the playlist beginning. The similarity between artists can be based on different measures, e.g., on their co-occurrence in publicly shared playlists. Since the CAGH method not only led to competitive performance results in terms of IR measures, but also led to a very good quality perception in a user study presented in (Kamehkhosh and Jannach, 2017), we also rely on a very similar recommendation approach in this work.

Commercial companies only sometimes reveal how parts of their machinery works. In a number of public presentations, Spotify, for example, revealed that they use matrix factorization based on implicit feedback for their "Discovery" feature (Johnson, 2014; Johnson and Newett, 2014). Little technical detail is however known about their "Radio" or playlist recommendations feature. After Spotify acquired The Echo Nest in 2014, a music intelligence platform that focused on the analysis of audio content, it was announced that they were planning to also utilize content-based techniques. In more recent presentations, such as (Steck et al., 2015), the authors report that Spotify uses an ensemble of different techniques including NLP models and Recurrent Neural Networks as well as explicit feedback signals (e.g., thumbs-up / thumbs-down), and also audio features for certain recommendation tasks. It remains, however, unclear from the presentations which techniques are used for which types of recommendations (radios, weekly recommendations, playlists, etc.).

According to the documentation of the Web API of Spotify, recommendations are aimed to create a *playlist-style* listening experience based on seed artists, tracks and genres. In this context, the available information for a given seed item is used to find similar artists and tracks. An analysis of Spotify's playlist continuation API revealed that it seems to have a tendency to recommend less popular items than some academic approaches (Jannach et al., 2016a). Since the above-mentioned CAGH method by design has a tendency to recommend comparably popular tracks, we decided to use Spotify's recommendation API as an additional source for track suggestions in our study.

In general, remember that our work is not aiming to compare different algorithms or advance the field in terms of prediction accuracy. Our study aims at understanding to what extent users adopt automated recommendations and how they influence their playlist creation. Therefore, we used algorithms that

have either shown to be generally well-perceived through a user study (CAGH) or were presumably optimized over years of practical deployment (Spotify API).

2.2 User Studies on Music Recommenders

Compared to the number of "offline" experiments based on historical data like listening logs or public playlists, the number of user studies on music recommenders is low.

Visualization and User Interaction Aspects. The probably most widely investigated aspect that has been explored through user studies relates to *visualization* aspects (Lehtiniemi and Ojala, 2013; Andjelkovic et al., 2016; Zhang and Liu, 2017; Andjelkovic et al., 2018) and to the design of the *user interface* (Baur et al., 2010, 2011; Bostandjiev et al., 2012; Kamalzadeh et al., 2016; Jin et al., 2017). Andjelkovic et al. (2016, 2018), for instance, introduced "MoodPlay" as an interactive music-artists recommender system, which integrates content and mood-based filtering functionalities in the user interface. Andjelkovic et al. (2018) conducted a user study involving 279 participants in which four versions of the interface were evaluated with various visualizations and interactions. The results showed, among other aspects, that the interface design and a certain combination of interactive features improved objective and perceived recommendation accuracy, as well as self-reported user satisfaction. In a similar approach, Zhang and Liu (2017) evaluated the usability of three alternative visualizations of a user's listening history (via Bean plot, Instrument plot, and Transitional Pie plot) through different user studies. Their results showed that the proposed visualization is useful for non-expert users for visual analysis tasks.

In another work, Baur et al. (2010) proposed an interactive track recommendation service called "Rush", which was optimized for touch-screen devices. Among other aspects, the authors analyzed its usability for left-handed and right-handed users. Furthermore, in (Baur et al., 2011), the authors introduced "Rush 2" as a mobile interface for playlist creation with varying levels of automation (from manual to automatic). A two-week diary study was conducted to evaluate the usability of the application and the satisfaction of the users. Other related user studies in this context, such as (Bostandjiev et al., 2012; Kamalzadeh et al., 2016) and (Jin et al., 2017), highlighted that explanations, visual representations, and user control mechanisms can be suitable means to increase user satisfaction.[1]

The visual design of a music recommender, the provided functionalities, its position on the screen, and its ease of use are important factors that influence its adoption by the users, see, e.g., (Knijnenburg et al., 2012). In our study, the design of the recommender is on the one hand inspired by the layout of

[1] For a recent review on user interaction aspects for recommender systems, see (Jugovac and Jannach, 2017).

existing tools, including the one by Spotify, where the recommended tracks are updated after each user interaction and where individual tracks can be added to the playlist under construction via drag and drop. Based on the insights from the literature and a pre-study described in Section 2.3, we furthermore decided to apply a grid layout for the recommendations, which has shown to be preferable in comparison to, e.g., one-dimensional list interfaces (Chen and Pu, 2010; Kammerer and Gerjets, 2010).

Immediate Consideration of User Feedback. In which ways users are enabled to state their preferences and how these preferences are integrated into the recommendation process is another area of research, which has been investigated to some extent through user studies (Pontello et al., 2017; Yakura et al., 2018). Pontello et al. (2017), for instance, presented a framework that relies on real-time user feedback to support personalized navigation in large media collections. To evaluate the proposed framework, they developed a Web application called "Mixtape" where users could provide a start track and were then presented one suggestion after each other. For each track, the users could provide a "like" statement or skip the track. One main result after analyzing over 2,000 navigation sessions was that the resulting playlists based on immediate feedback were more coherent in terms of the artists in comparison with tag-based and artist-based navigation approaches.

In a more recent work, Yakura et al. (2018) proposed a recommender system called "FocusMusicRecommender" for playing music while working. Among other functionalities, they introduced a "keep listening" feedback button that together with the "skip" button enables the system to determine songs that users like and dislike. The authors also estimated the concentration level of users based on various indicators such as keyboard input, mouse input, and Web browsing activity. The estimated concentration level was then adopted to refine the users' preference level. The value of considering the user's concentration level was then validated in a within-subjects user study, where the participants reported on their impressions regarding the different tracks that were played by the recommenders. The study revealed that considering the estimated concentration level was useful, leading to a selection of tracks that is more suitable during work.

Differently from the systems proposed by Pontello et al. (2017) and Yakura et al. (2018), the users in our study can not explicitly like, dislike, or remove one of the provided recommendations. However, each addition of a recommendation to the playlist under construction is considered as a positive implicit feedback signal, and similar to the work by Pontello et al. (2017), this feedback is immediately taken into account to compute the next set of recommendations.

Quality Perception of Recommendations. Another group of user studies in the literature explore the quality perception of music recommendations (Sinha and Swearingen, 2001; Barrington et al., 2009; Jones, 2010; Kamehkhosh and Jannach, 2017). In (Kamehkhosh and Jannach, 2017), the results of a study

involving 277 subjects[2] indicated that users generally preferred recommendations that were coherent with the recently played tracks in different dimensions. The results also revealed that the participants tended to evaluate recommendations better when they already knew the track or the artist. In a similar work, Barrington et al. (2009) compared different music playlisting approaches in a user study involving 185 participants. The compared methods included Apple iTunes' Genius collaborative filtering based system, a method based on artist similarity, and one based on the similarity of acoustic content. Their results indicated that, among other aspects, the general similarity of the recommendations with the seed track had a major importance, especially in terms of sound and style, and that showing artist and song names had a large influence on how users evaluated the generated playlists of each algorithm.[3]

In another work, Jones (2010) reported the results of two within-subjects user studies to compare the user acceptance and adoption of the recommendations provided on two popular music websites, *Pandora.com* and *Last.fm*. The participants of the studies were asked to register for both services and were queried about their preferences after using the services. The results indicated that users strongly preferred Pandora over Last.fm, mainly because of its perceived superiority in terms of recommendation quality, but also in terms of novelty and enjoyability of recommendations as well as its user interface. Last.fm and Pandora mainly provide recommendations through Web radios, and users of these services are not provided with additional recommendation lists from which they can actively pick individual tracks as they can do in our study. An interesting observation in the context of their study is that the participants were not particularly interested to own or purchase the recommended tracks of either service. Only between 2 % to 4 % of the users answered positively to the corresponding questionnaire items "I would like to own the recommended songs" and "I would purchase the recommended songs given the opportunity."

In a similar work, Lee and Waterman (2012) studied the preferences of users regarding severak commercial music streaming services, including Pandora and Last.fm. Like in the study by Jones (2010), the participants generally preferred Pandora over Last.fm. Furthermore, the study indicated that users tend to prefer services that are more suited for discovery, with 12 % of the participants selecting discovery of new music and artists as the main purpose of the recommendations. More recently, Mäntymäki and Islam (2015) performed a user study involving 374 Spotify users with the goal of identifying the main factors for the continued use of the service. One of their main findings was that discovery is a driving factor for renewing *premium* subscriptions. For users of the free version, discovery was, however, less relevant.

[2] The fact that the study of Andjelkovic et al. (2018), Kamehkhosh and Jannach (2017) and the current study all have about 270 participants is a coincidence.

[3] Note that the authors in (Barrington et al., 2009) pre-selected 12,000 songs from which iTunes' Genius built the playlists for the experiment. It is, therefore, not clear whether or not the results would still be valid in music streaming services with much larger music libraries.

Generally, in most of the works discussed here, the main focus was on comparing the quality perception of different competing algorithms. In an alternative approach, Sinha and Swearingen (2001) examined the perceived quality of six different online recommender systems by comparing their recommendations with recommendations that were made by the friends of a user. Their results showed that users perceived the recommended items by recommender systems often "new" and "unexpected", while recommendations of friends mostly served as reminders of what users liked in the past. Similar to this work, our main focus is not on the optimization of a certain recommendation algorithm, but on the adoption of recommendations in general and their effects on users.

Choice of Tracks for Playlists in General. Finally, a number of studies in the literature are concerned with understanding the factors that influence which tracks users choose for inclusion in a playlist in different situations (Pauws, 2002; Swearingen and Sinha, 2002; Cunningham et al., 2006, 2007; Lehtiniemi, 2008; Lamont and Webb, 2011; Lee et al., 2011; Stumpf and Muscroft, 2011) and (Kamalzadeh et al., 2012). For example, the results of an early user study conducted by Pauws (2002) showed that playlists containing personalized tracks that were selected automatically for particular contexts ("listening to soft music" and "listening to lively music") were preferred by users over randomly assembled playlists. Lehtiniemi (2008) also conducted a user study with 42 participants to evaluate the performance of context-aware recommendations in comparison to random recommendations. Their results led to very high levels of user satisfaction with the proposed context-aware approach (called "SuperMusic"), which highlights the importance of considering the user's context when selecting tracks for playlists. In another work, Lamont and Webb (2011) explored short-term and long-term factors that influence the users' music preferences from a psychological point of view. The authors conducted a diary study for one month with nine undergraduate students with follow-up interviews with two participants. Their results indicated that musical favourites are subject to rapid change and that they are highly context-dependent, as was observed also in the aforementioned studies.

In (Cunningham et al., 2006), both interviews and Web forum posts related to the construction of playlists were analyzed. The authors found that playlists are often created with a theme or topic in mind, e.g., a genre, an event or a mood. Similarly, also the study in (Swearingen and Sinha, 2002) showed that mood, genre, and artists are the most important factors for users when selecting the tracks, which is in line with the outcomes of the studies of (Cunningham et al., 2006; Stumpf and Muscroft, 2011) and (Kamalzadeh et al., 2012). Moreover, the user study from (Lee et al., 2011), in which the participants were asked to evaluate playlists that were automatically created based on a seed track, showed that factors such as variety, metadata, personal preferences, familiarity, and the combination of familiar and new music also strongly affect the users' perception of playlist quality.

Finally, Kjus (2016) analyzed the usage of playlists curated by WiMP (now renamed to TIDAL) during summer 2012. Their investigations revealed that playlists that promoted new music tended to have high but short usage peaks shortly after being released. Historical compilations and contextualized playlists, in contrast, had a much longer life span. Interviews with a focus group furthermore revealed that music is mainly discovered through friends and acquaintances, although streaming platforms were often used subsequently to explore the discovered music more deeply. The interviews also indicated some form of distrust towards both the curators (especially when the participants felt that some tracks were included for commercial reasons) and algorithms (especially when their accuracy was too low). In addition, it turned out that users tend to feel overwhelmed by the quantity of available tracks.

In our user study, we asked the participants to create a playlist for one of several given topics, and explicitly asked them about their selection rationale after the task. Differently from some previous studies, we therefore rely also on explicit user statements about the choice process and do not aim to reconstruct the users' motivations solely from the resulting playlists.

2.3 Observations from a Pre-Study

In (Kamehkhosh et al., 2018), we discussed the results of a preliminary user study that was executed to inform the design of the study reported in this paper. The general design of this pre-study is very similar to the one presented here. The task of the participants was to create a playlist for one of several predefined topics, and one part of the participants was supported by a recommendation system.

Based on the observations and insights from this previous study, we improved the design of the new study in different ways. First, in the pre-study, we relied solely on Spotify's API to generate the next track recommendations. Since it is unknown how Spotify generates the recommendations, we created the recommendations in the new study using two sources as mentioned above. Including the academic recommender therefore allows us to move away from pure "black-box" recommendations. In addition, we also included further questions in the post-task questionnaire that are designed to help us understand the reasons why participants selected certain tracks from the provided recommendations.

Another interesting observation in our previous study was that there was a subset of participants who never picked any of the recommendations, but whose track choices were apparently *influenced* by the provided recommendations.[4] This phenomenon manifested itself in particular in the choice of the artists. In our previous study, we had not anticipated such an effect. In the new study, we include a corresponding measurement and explicitly ask the participants in the post-task questionnaire about this potential influence.

[4] See (Köcher et al., 2018) for a recent study of such effects in the e-commerce domain.

2.4 Discussion

The study presented in this paper is inspired by insights from existing work in two main ways. First, we relied on existing works to inform the design of the recommendation service to be used in our study. Given the abundance of existing research on that topic, we decided to combine two complementary algorithms. The first one, the CAGH algorithm, is a simple yet particularly competitive algorithm and has a tendency to recommend relatively popular tracks. The second one, Spotify's playlist continuation algorithm, furthermore allows us to rely on industry-strength recommendation technology. Their algorithm, as discussed above, also has a tendency to recommend less popular tracks. Besides the recommendation algorithm, the user interface shows to have a significant effect on user satisfaction. Following insights from the literature, we adopted a grid layout in our study, which was often found to be preferred over one-dimensional list interfaces. Furthermore, the recommendation system used in our study supports the incorporation of user actions in real-time, leading to immediate updates when items are added or removed from the playlist.

The second inspiration from existing research relates to the problem of identifying the factors that influence the selection of the tracks and the adoption of the recommendations. Several factors were found in the literature to have a potential influence on the users' behavior. These factors for example include the similarity and diversity of the playlist in terms of the topic, theme, mood, genre or artists. Furthermore, the general user preferences and the users' familiarity with certain tracks or artists are of course relevant as well. However, existing studies are often limited to either the quality perception of different recommendation approaches (e.g., content-based versus collaborative filtering), or the analysis of the characteristics of the created playlists. In this present study, we aim at a more comprehensive approach. We both compare the behavior of users when supported with recommendations and without recommendation support, we analyze the resulting playlists in different dimensions, and we investigate further aspects of the users' choice process and their satisfaction through a questionnaire.

3 Research Questions

Our research questions revolve around the problem of understanding in which ways system-provided recommendations have an impact on the behavior of users when creating playlists. In that context, we consider a *playlist* to be a sequence of tracks that is created manually for a given purpose, e.g., for listening during a road trip. In the research literature, alternative types of playlists exist as well, e.g., *club playlists*, which are made by DJs in clubs, or *personalized radio playlists*, which are generated by Web music services like Spotify, see (Bonnin and Jannach, 2014). In particular, the latter type of auto-generated playlists is not in the focus of our work, as we are interested in the

usefulness of a recommender system as an assistive tool during manual playlist construction.

Regarding our research questions, existing works on music recommendation often focus on the improvement of the prediction accuracy of the underlying algorithms. However, as discussed above, high prediction accuracy does not necessarily correlate with a high adoption of the recommendations (Jones, 2010), and more research is required to understand to what extent users actually rely on such a recommendation service in the first place. If we observe a high adoption rate, this would support the assumption that such a recommendation service is relevant and provides potential value for the consumer. Low adoption, in contrast, could indicate that today's technology is not yet at a state where it is considered particularly helpful. This would mean that further studies are required to understand why users do not rely on the system's recommendations. Our first goal is therefore to assess to what extent recommendations are adopted during playlist construction.

Research Question 1 (RQ 1) *How high is the adoption rate of recommendations during playlist creation?*

Independently of RQ 1, we are interested in the reasons why users select certain tracks from the provided recommendations and not others, or why they do not use any recommendations at all. Although the match with user preferences is commonly considered a key factor for the adoption of recommendations, previous studies have shown that optimizing for prediction accuracy alone might not be sufficient. Other factors can play an important role as well and a better understanding of the relative importance of these factors will help service providers design better-accepted systems. In our study, we are specifically interested in three types of factors, which we will discuss in more depth below: *user-related* factors, *system-related* factors and *music-related* factors.[5]

As *user-related* factors, we consider the match of the recommendations with the users' general taste, the users' expertise and enthusiasm for music and their attention.

Research Question 2 (RQ 2) *How important is each user-related factor (i.e., the match with user preferences, the expertise and enthusiasm of the users and their attention towards the recommendations) for the adoption of recommendations?*

The *system-related* factors considered in our study include the trust the users have in the system, the perceived ease of use and the helpfulness of the system.

Research Question 3 (RQ 3) *How important is each of the system-related factors (i.e., the trust in the recommender system, the perceived ease of use, and the helpfulness of the recommender) for the adoption of recommendations?*

[5] The used categorization follows a common perspective in the field of recommender systems, where various aspects regarding user, item, and system characteristics have to be considered in parallel to optimize the overall recommendation service (Xiao and Benbasat, 2007; Konstan and Riedl, 2012; Jannach et al., 2016b).

Music-related factors relate to (i) general quality criteria of playlists and to (ii) the characteristics of the provided recommendations. Regarding *generally desirable* playlist characteristics, participants of previous studies often state that criteria such as the diversity of the artists or the homogeneity of the track characteristics are of major importance to them (Cunningham et al., 2006; Fields, 2011; Lee et al., 2011). However, it is often not fully clear from these studies to what extent the users actually consider these subjective criteria when they decide to accept a recommendation.

The characteristics of the recommendations themselves can also impact the adoption of a recommendation. For instance, previous research provided indication that the adoption is dependent on the diversity of the recommendations (Jones, 2010; Castagnos et al., 2013).

Research Question 4 (RQ 4) *How important is each of the music-related factors (i.e., the quality criteria of the playlists and the characteristics of the recommendations) for the adoption of recommendations?*

Generally, music can be played in various contexts, and depending on these contexts, the individual quality factors might be of different importance. Schedl et al. (2018) differentiate between the *listening context*, for instance the location or activity, and the *listening purpose*, for instance mood regulation or social bonding. We consider these two types of context in our study by asking participants to create a playlist for a certain listening context (e.g., while driving a car) and a certain purpose (e.g., to calm down). We use the term *playlist topic* to denote this combination of listening context and purpose.[6]

The distinction between the listening context and purpose was also studied in the field of music sociology in form of the *functions of music* in everyday life. For example, interviews conducted by DeNora (2000) showed that the interviewees use music in their everyday life – in contexts such as shopping malls or exercise classes – to organize their internal and social world. Similarly, North et al. (2004) conducted a user study with 346 people about the music had heard during the day. Their results also highlighted the relevance of context and purpose for the users' perception of music. For instance, they concluded that people like music that they listen in isolation more than what they listen in the presence of others.

Several other works have pointed out the general importance of contextual factors when providing recommendations (Pichl et al., 2015; Dias et al., 2017; L'Huillier et al., 2017). However, the importance of contextual aspects in comparison to other quality factors has, to the best of our knowledge, not been studied so far. Our research question therefore is as follows.

Research Question 5 (RQ 5) *To what extent does the adoption of recommendations depend on the topic (context and purpose) of the playlist to be created?*

[6] Note that both terms "topic" and "theme" are usually used interchangeably in the literature for the same concept. For instance, Hariri et al. (2012) use the term "topic" while Cunningham et al. (2006) use the term "theme." In this paper, we mainly use the first one, although we consider both interchangeable.

Although the goal of recommender systems is generally to recommend items that the user may like or need, they can also influence the choices of the users and make them consume items they would not have consumed otherwise, a phenomenon often referred to as persuasiveness (Zanker et al., 2006; Yoo et al., 2012). Persuasiveness is a multi-faceted concept, for which a number of definitions can be proposed (O'Keefe, 2002). In this paper, we rely on the definition of persuasive technologies proposed by Fogg (2003): "interactive computing systems designed to change people's attitude and behaviors."

Persuasiveness has previously been observed for various types of recommendations including, TV shows (Adomavicius et al., 2011), traveling (Gretzel and Fesenmaier, 2006), etc., but to the best of our knowledge, was never studied for music recommendation. Recent research has shown that even the mere presence of recommendations can have a significant influence on the choices of the users (Köcher et al., 2018, 2016), i.e., users tend to choose items that share some characteristics with the recommended items. The provided recommendations may thus be inspiring for users and help them find relevant tracks even if the recommendations are not directly used.

Research Question 6 (RQ 6) *To what extent do the recommendations influence the choices of the users?*

4 Study Design

To answer the research questions, we conducted an online study. All participants were asked to create a playlist for a predefined topic using an online application that was developed for the purpose of the study.The online experiment consisted of the following three main parts.

Intro – Select a Topic. In the first step – on the welcome page of the application – the participants were introduced to the task and to the terms and conditions of the experiment. All participants were then asked to select a topic for the playlist to be created from the following predefined list (see also Figure 1):

1. *Chillout – Crazy songs.* "Crazy tracks to let it go, let the storm rage on for an evening at home after a stressful workday."
2. *Chillout – Relaxing tracks.* "Relaxing tracks for an evening chillout at home after a stressful workday."
3. *Roadtrip – Bank robbery.* "Tracks to get energized while escaping after a bank robbery on a summer day."
4. *Roadtrip – Singing along.* "Tracks to sing along on a roadtrip on a summer day."
5. *Work – Motivation.* "Tracks to get motivation to work when bored at office."
6. *Work – Focusing.* "Tracks to better focus on some complex topic at office."

The main reason for asking the participants to choose such a topic was to answer RQ 5 about how topics and adoption factors are related. We selected

the topics as follows. We initially defined 20 topics by combining different contexts with different purposes. We then presented these topics to a group of about 60 students and asked them to select those for which they felt capable to create a playlist in a few minutes. We then selected the three most frequently selected contexts together with their corresponding purposes.

For each of the finally selected six topics, we provided a brief explanation to make the *context* as well as the *purpose* of the playlist clear. To avoid *order effects*, the order of the topics was randomized across all participants.

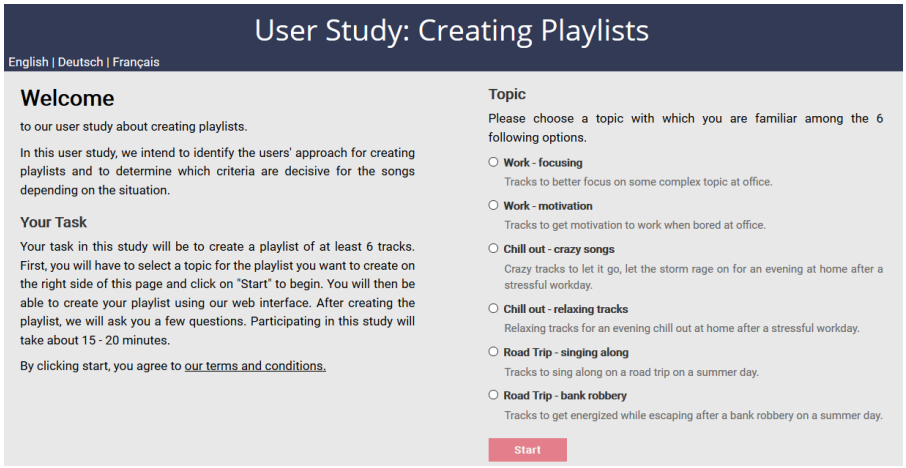


Fig. 1: First part of the user study – introduction and list of topics.

Main Task – Create a Playlist. After selecting a topic, the participants were forwarded to the playlist-creation page. A short instruction on how to use the playlist creation tool was presented to the participants at the beginning. In order to answer RQ 6, i.e., how do the recommendations influence the choices of the users, the participants were automatically assigned to one of two groups: one group with recommendations support (called `Rec`) and one group without recommendation support (called `NoRec`). Both groups could add tracks to their playlist using an advanced search interface. The treatment group (`Rec`) additionally received recommendations at the bottom of the page during the playlist creation task (see Figure 2). The control group (`NoRec`) was shown the same interface but without the recommendation section at the bottom, i.e., these participants could only use the search functionality. The assignment of the trials was based on a round-robin scheme across the participants to obtain roughly the same number of participants for each group.

To help participants start the playlist creation process, one song was suggested to the participants of both groups. They could either accept or reject it. The suggested start song of each topic was selected from a topic-related popular playlist on Spotify.

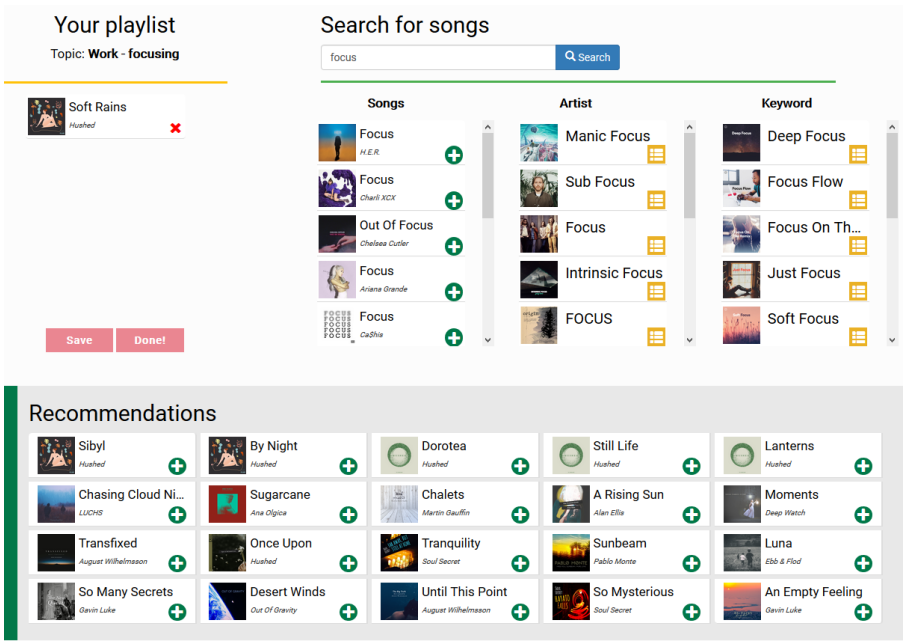


Fig. 2: Playlist creation tool used in the study.

The lists of recommendations of the `Rec` group contained 20 tracks. The recommendations were displayed after the first track was added to the playlist. The list was then updated automatically every time an item was added to the playlist or was removed from it. Ten recommendations of the list were retrieved through the recommendation API of Spotify. Although it was possible to set target values for some attributes such as energy, loudness, or popularity over the API, we used the default parameter values to avoid any bias.[7] Note that the recommendation API accepts up to 5 seed tracks. We, therefore, provided the last 5 tracks of the playlist to retrieve recommendations.

The other 10 recommendations were generated by an artist-based approach we called "Related-Artists Greatest-Hits" or RAGH. Similar to the CAGH algorithm (Bonnin and Jannach, 2014), the RAGH method recommends the greatest hits of artists that appeared in the given playlist and of artists that are "related" to the artists of that playlist. To determine the greatest hits as well as the related artists, we relied on Spotify's API.[8] The RAGH recommendations were finally generated as follows. First, we determined the single greatest hit of each artist of the last 5 tracks of the given playlist. The remaining places of

[7] For more details about request parameters and tunable track attributes of Spotify's recommendation API, see `https://developer.spotify.com/documentation/web-api/reference/browse/get-recommendations/`.

[8] According to the documentation of the Web API of Spotify, related artists are determined based on analysis of the listening histories of Spotify users.

the 10-item recommendation list of RAGH were filled with the greatest hits of their most related artists.

These two lists – the RAGH-based one and the one based on the Spotify API – were then merged into one recommendation list as follows. To avoid order effects, each time the recommendation list was updated, the top-10 recommendations of one of the recommenders (Spotify or RAGH) were randomly selected to be included first in the recommendation list. The top-10 recommendations of the other recommender were then placed at the end of the recommendation list as the 11^{th} to the 20^{th} recommendation.

When the participants created their playlists, they were able to listen to 30-second previews of the searched and recommended tracks. The previews were also obtained via Spotify's API.[9] The order of the tracks in the playlist could be changed via drag-and-drop.

Once the playlist contained at least six tracks, the participants could proceed to the post-task questionnaire. One reason for selecting six as the minimum number of tracks in a playlist relates to the fact that Spotify accepts up to five seed tracks. Moreover, with six tracks the complexity (cognitive load) of the playlist creation task should remain with manageable limits for the participants (Shiffrin and Nosofsky, 1994). At the same time, requiring at least six tracks ensures that the created playlists are not too short. During the process, the participants were also given the opportunity to save their playlist for later use as a CSV file which contains the track names and their Spotify identifiers.

Post-Task Questionnaire. In the last step of the user study, participants were asked a series of questions concerning their created playlist and, if applicable, concerning the provided recommendations. The questionnaire items are partially adapted from (Knijnenburg et al., 2012). In particular, these questions should help us find answers to the above-mentioned research questions and help us understand additional aspects and relationships related to the research questions. One part of the questionnaire items was presented to all participants. Another set of additional questions was shown to the `Rec` group who had received recommendations during the playlist creation task.

Questionnaire items for all participants. To understand if recommendation support had a positive impact on the users' *satisfaction* with their created playlists, we asked all participants to what extent they believed that their chosen tracks fit the topic, if they liked their chosen tracks, if they liked their created playlist, and if they were confident enough about their playlist that they were ready to share it. The specific questions are presented in List 1 of Table 1.

Furthermore, we asked all participants about the *perceived difficulty* of the playlist creation task (Table 1 – List 2). Participants could express their agreement with the provided statements on a 7-point Likert scale item. Their

[9] The excerpts were usually not the first 30 seconds of the tracks.

List 1. How satisfied are you with your chosen tracks?
Item 1. My chosen tracks fit my preferences. (7-point Likert item)
Item 2. My selected tracks fit the chosen topic. (7-point Likert item)
Item 3. I like the tracks I have chosen. (7-point Likert item)
Item 4. I like my created playlist. (7-point Likert item)
Item 5. I am satisfied with the transition between the tracks I have selected. (7-point Likert item)
Item 6. I am confident enough about my created playlist that I am ready to share it. (7-point Likert item)
List 2. Perceived difficulty of the playlist creation task.
Item 7. Overall, it was a difficult task for me to create this playlist. (7-point Likert item)
List 3. Quality criteria for playlists.
Item 8. Which of the following characteristics were relevant for your playlist? (Two drag-and-drop lists as displayed in Figure 3) *Lyrics*: The lyrics should fit the topic. *Order*: The songs should be in a certain order. *Transitions*: The beginning of each track should not be too different from the ending of the previous track. *Musical Characteristics*: The musical characteristics of the songs should be homogeneous. Examples for this are: Tempo, Loudness, Energy. *Variety*: The songs should be diverse in terms of artists. *Popularity*: The songs should be known and generally popular. *Freshness*: The songs should be new/up-to-date.
List 4. Which of the following statements apply to you?
Item 9. I am a music enthusiast. (7-point Likert item)
Item 10. Compared to my peers, I listen to a lot of music. (7-point Likert item)
Item 11. I create playlists a lot. (7-point Likert item)
Item 12. I only use shared playlists on music platforms like Spotify. (7-point Likert item)
List 5. Personal information
Item 13. Age group (drop-down menu: under 20, 20–30, 30–40, 40–50, above 50)
Item 14. Comments, suggestions or considerations about the recommendations, playlist or the study itself (free text)
Item 15. Email address (optional; text box)

Table 1: Questionnaire items for all participants.

answers should help us analyze the *perceived ease of use* and its impact on the adoption of recommendations, in particular to address RQ 3.

In addition, all participants were presented with a list of *quality factors* for playlists that were mentioned in the literature. The features were either related to individual tracks (e.g., popularity or freshness) or to the list as a whole (e.g., artist homogeneity), see Table 1 – List 3. The participants were asked to rank these quality factors by decreasing order of relevance for their playlist, or mark them as irrelevant (see Figure 3). The rankings provided by the participants

Which of the following statements apply to the recommendations shown to you?
Diversity
Item 1. The recommendation lists were varied. (7-point Likert item)
Item 2. The recommendation lists included tracks of many different genres. (7-point Likert item)
Item 3. The recommendation lists included tracks of many different artists. (7-point Likert item)
Item 4. All recommended tracks seemed generally all similar. (7-point Likert item)
Discovery
Item 5. The recommender provided interesting tracks that I did not know. (7-point Likert item)
Item 6. The recommender provided interesting artists that I did not know. (7-point Likert item)
item 7. I knew most of the recommended tracks. (7-point Likert item)
Item 8. I knew most of the recommended artists. (7-point Likert item)
Effectiveness
Item 9. The recommender was generally useful. (7-point Likert item)
Item 10. I could save some time using the recommender. (7-point Likert item)
Item 11. The recommendations matched my interests. (7-point Likert item)
Item 12. The recommendations fit the given topic. (7-point Likert item)
Difficulty
Item 13. Recommendations made the task of making decisions overwhelming. (7-point Likert item)
Item 14. The recommender gave too many recommendations. (7-point Likert item)
Item 15. I liked the layout of the recommendations. (7-point Likert item)
Popularity and freshness
Item 16. The recommended tracks were generally popular. (7-point Likert item)
Item 17. The recommendations were new and up-to-date. (7-point Likert item)

Table 2: Additional questionnaire items for the `RecUsed` and `RecSeenNotUsed` groups.

allow us to contrast their expressed quality criteria with the characteristics of the adopted recommendations, which should help us answering RQ 4.

Finally, to investigate if users behave differently depending on demographics and music expertise, and to answer RQ 2, we asked all participants a number of questions about their age, music enthusiasm and how often they listen to music and create playlists. These latter questions were asked to estimate the expertise of the participants. The specific items are presented in List 4 and 5 of Table 1.

Additional questions for the participants with recommendation support. The participants of the `Rec` group were asked to answer additional questions about the recommendations that they received during the playlist creation task, see Table 2. To be sure that only those participants who had actually looked at the recommendations get these questions, we first asked them in which ways

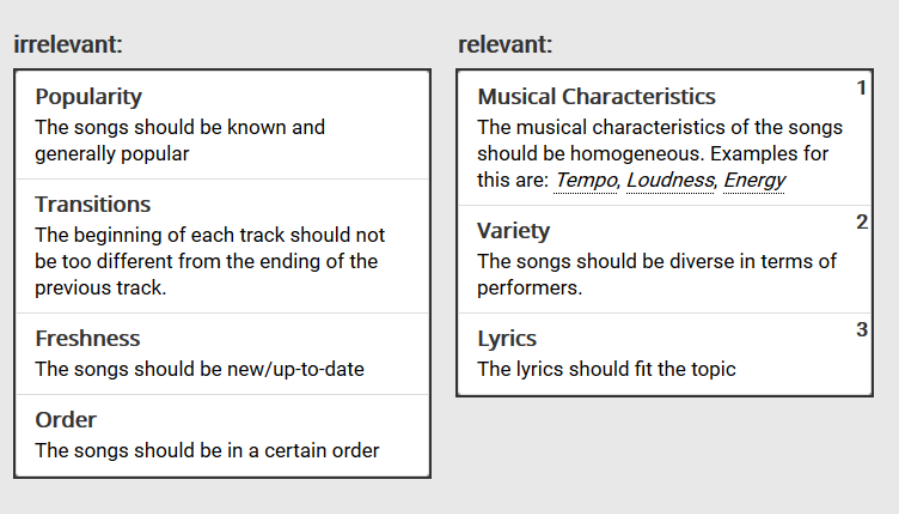


Fig. 3: A list of quality factors was presented to the participants via drag-and-drop-enabled lists. The predefined list of quality factors was presented in the left list in randomized order. The right list was initially empty. Participants could move factors that were relevant to them to the right list. The order of the elements in the right list should be used to express the relative importance of the factors.

they had used the recommendations. Specifically, our goal was to distinguish between the following types of participants. Those who did not notice the recommendations at all (called `RecNotNoticed`), those who noticed the recommendations but did not pay attention to them (`RecNoAttention`), those who looked at the recommendations but did not use them (`RecSeenNotUsed`), and finally those who actually used the recommendations (`RecUsed`) (see Figure 4).

Accordingly, we asked the subgroups of the `Rec` group who actually used the recommendations (`RecUsed`) and the ones that looked at the recommendations but did not use them (`RecSeenNotUsed`) about the quality of the recommendations and the recommender system in general. Again, participants could express their agreement with the provided statements, e.g., "The recommendation lists were varied.", on a 7-point Likert scale. The answers to these questions correspond to subjective ratings of various quality criteria of the provided recommendations. This will allow us to look for possible correlations between the perceived quality of the recommendations and their adoption, which is the main focus of RQ 4.

To obtain a better understanding of why some of the users who received recommendations did not use them, we asked the respective participants some further questions. For example, we use the responses of the participants of the `RecSeenNotUsed` group to the statement "Although I did not use the

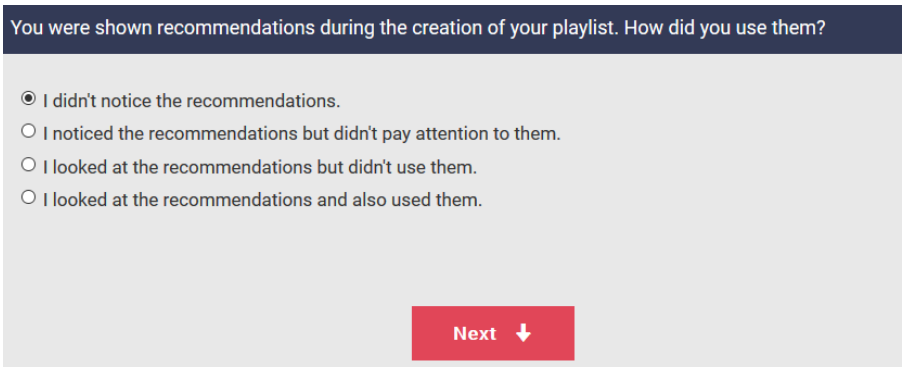


Fig. 4: We asked the participants how they had used the recommendations during the playlist creation task.

recommendations, they influenced my selected tracks." to address RQ 6 about the influence of the recommendations on the users' choices. These questions are listed in Table 3.

5 Results

Overall, 270 participants completed the study[10]; most of them were university students from Germany, who were recruited via invitations sent to university mailing lists; a smaller part was recruited via invitations on social network sites. Most (88 %) of the participants were aged between 20 and 40.[11]

On a scale from 1 and 7, the median of the self-reported enthusiasm for music was 6 and the median of the self-reported values on how often they listen to music was 5, i.e., the majority of the participants considered themselves experienced or interested in music. Most of the participants, however, do not create playlists regularly; only 34 % of the participants responded to the corresponding statement ("I create playlists a lot.") with a 5 or higher (median=3).[12]

It is worth noting that generally the participants presented a high level of task engagement. They spent on average 7.46 minutes on the playlist creation task and created playlists that contained on average 8.98 tracks although the requested size was only 6. The participants were also generally satisfied with their created playlists. In fact, more than 90 % of the participants – independent of the treatment group – responded to the statement "My chosen tracks

[10] Note that these were other participants than those that took part in our pre-study.

[11] As we asked for the age group of the participants and not their exact age, see Table 1, List 5, we can only use the lower bound (21 years old) and the upper bound (31 years old) when computing the average age of the participants.

[12] The collected data is ordinal, i.e., a ranking of the response levels is possible. However, we cannot assume equidistance between the response levels, and reporting mean and standard deviation values in such a case is often considered questionable in the literature.

Items shown to the participants of the `RecNotNoticed` group
Item 1. Any reason for not noticing the recommendations? (free text)
Item 2. Would you have considered the recommendations if you had noticed them? (7-point Likert item)
Items shown to the participants of the `RecNoAttention` group
Item 1. I'm not convinced by the potential of recommender systems. (7-point Likert item)
Item 2. I do not trust recommender systems. (7-point Likert item)
Item 3. I do not feel at ease with the use of recommender systems. (7-point Likert item)
Item 4. I can find better tracks without the help of recommender systems. (7-point Likert item)
Item 5. Any other reason for not paying attention to the recommendations? (free text)
Items shown to the participants of the `RecSeenNotUsed` group
Item 1. I did not like any of the recommendations in the lists. (7-point Likert item)
Item 2. I do not trust recommender systems. (7-point Likert item)
Item 3. I do not feel at ease with the use of recommender systems. (7-point Likert item)
Item 4. I only use shared playlists on music platforms like Spotify. (7-point Likert item)
Item 5. Although I did not use the recommendations, they influenced my selected tracks. (7-point Likert item)
Item 6. Any other reason for not using the recommendations? (free text)
Item shown to the participants of the `NoRec` group
Item 1. I would have liked to get automated recommendations during the creation of my playlist. (7-point Likert item)

Table 3: Questionnaire items for the participants with no recommendation support or those who did not use the recommendation functionality.

fit my preferences" with a rating of 5 or higher. Moreover, the majority of them stated that they were confident to share their playlists – the median value to the corresponding statement is 5. About 58 % of them saved their created playlists, which indicates a certain intention of actually reusing them. These generally high engagement and satisfaction levels make us confident that the users accomplished the tasks thoroughly.[13]

[13] Note that we did not aim to study the effect of demographic factors in this study. We, therefore, did not ask about the gender of the participants. However, looking at the participants' email addresses shows that among 155 participants who entered their official university email addresses and where we were sure about the gender, the proportion of men and women was 122 (79 %) to 33 (21 %), respectively.

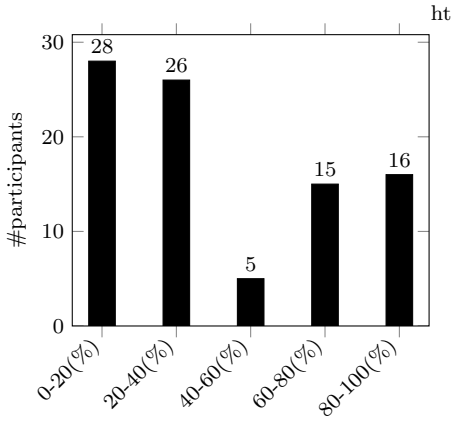


Fig. 5: Bin distribution of the proportion of the number of accepted recommendations to the total number of tracks in playlists (playlist sizes) for the participants of the `RecUsed` group.

RQ 1: Adoption of recommendations

To answer this research question, we first looked at the proportion of participants who were presented with recommendations and actually used them. As mentioned previously, half of the participants (135 participants) were assigned to the `Rec` group. As many as two thirds (67 %) of these participants drag-and-dropped at least one of the recommended tracks to their playlists. We denote this group as `RecUsed`. On average, each participant of the `RecUsed` group used 3.6 recommendations, which represents 39 % of the tracks of their playlists. This value is 2.4 recommendations per playlist when all participants of the `Rec` group are considered.

Overall, we interpret these numbers as strong indicators of the general usefulness of a recommendation component in the music domain. These findings are also in line with our preliminary study (Kamehkhosh et al., 2018), where the proportion of users who used at least one recommendation was about 50 %, and 38 % of the tracks in the playlists of these participants came from the recommendations.

To better understand if participants either mostly relied on the recommendations, mostly used the search interface, or if they often used both options, we counted the percentage of tracks that were taken from a recommendation for each playlist. The resulting distribution is shown in Figure 5. The results show that most participants decided to rely mostly on one of the two options. Only in a few (five) cases – represented by the bar in the middle of the figure – the participants used both options in parallel. In these cases, about half of the items were taken from the recommendations and half of the items were retrieved via the search interface. Overall, therefore, only a few participants used a balanced mix of the available functionality.

The numbers so far are based on the actual use of the recommendations. To obtain a more comprehensive picture, we additionally asked the participants of the `NoRec` group – those who had no recommendation support – whether they would have liked to get automated recommendations during the playlist cre-

ation task (cf. Table 3). In fact, about 55 % of these participants answered the question positively (i.e., assigned a rating greater or equal to 5), see Figure 6.

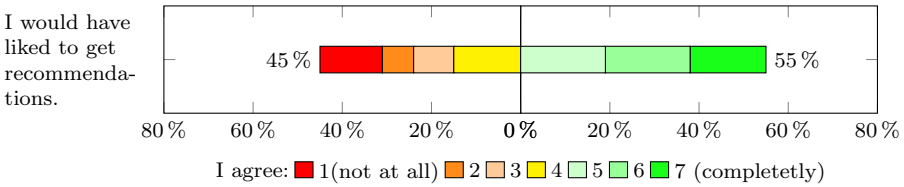


Fig. 6: Distribution of the responses of the participants of the `NoRec` group to whether they would have liked to get recommendations during the playlist creation task.

Overall, our results confirm that the participants generally liked the recommendation feature and actually used it to a quite large extent. Recommendation support during playlist creation can therefore be considered a feature of high potential value for customers of an online music service.

RQ 2: User-related factors for the adoption of recommendations

We now turn our attention to the different adoption factors and the effects of the recommendations on the behavior of the participants. We start our discussion with following user-related factors: the users' preferences, their expertise and enthusiasm for music, and their attention with respect to the recommendations.

Match with user preferences. According to the answers to our questionnaire, 40 % of the participants of the `RecSeenNotUsed` group did not adopt any recommendation because they did not like the recommended tracks (see the first bar in Figure 7). In other words, the fact that the recommender did not match well the preferences of the users explains less than half of the cases when no recommendation was adopted. The distribution of other reasons stated by the participants for not adopting recommendations are shown in Figure 7.

We also measured the correlation between the number of times the participants selected a track from the recommendations and the ratings they provided to the question about the match of recommendations with their preferences.[14] We obtained a correlation of 0.35 ($p < 0.001$), which is considered as a moderate correlation (Corder and Foreman, 2014). In other words, the match of recommendations with the users' preferences only moderately influenced the

[14] As the collected data about the quality of recommendations are ordinal (between 1 and 7), we used *Spearman*'s correlation measure, which is the nonparametric version of Pearson's correlation measure.

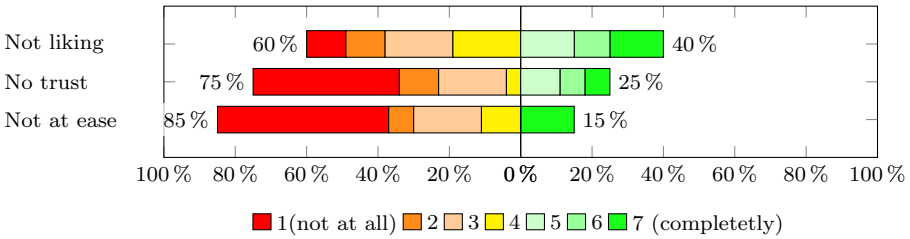


Fig. 7: Distribution of the responses of the participants of the `RecSeenNotUsed` group to why they would have not used any recommendations during the playlist creation task. "Not liking" corresponds to the statement "I did not like any of the recommendations in the lists.", "No trust" corresponds to the statement "I do not trust recommender systems.", and "Not at ease" corresponds to the statement "I do not feel at ease with the use of recommender systems."

adoption of the recommendations. Overall, these results are in line with previous studies showing that being able to predict the user's preferences with high accuracy is not fully sufficient to obtain a high adoption rate for the recommendations (Jones, 2010; Nilashi et al., 2016).

Expertise and enthusiasm of users. Looking at the answers to the questions related to the user expertise, we observed a strong difference between the group of participants who were very frequent playlist creators and the others.[15] The frequent playlist creators represent about 13 % of the participants. They adopted on average 1.5 recommendations, whereas other participants adopted 3 items on average. The difference is statistically significant ($p = 0.03$).[16]

Those participants who do not very frequently create playlists reported different levels of music enthusiasm. Again, we found differences in their behavior that seem to depend on their expertise. Participants with comparably low music enthusiasm – with a rating lower than 4 for the corresponding question – only adopted about 1.7 recommendations on average. Participants with higher enthusiasm in contrast picked 3.23 tracks. The difference is again statistically significant ($p = 0.03$).

Overall, these results show a high importance of the expertise and enthusiasm of the users, as the participants who did not adopt many recommendations were either experts in playlist creation or had low enthusiasm for music. As a result, service providers might want to implement additional means to particularly stimulate users with apparently low music enthusiasm to make more use of the provided recommendation functionality.

[15] We considered those as very frequent creators who answered with the highest rating to the corresponding questionnaire item.

[16] To test for statistical significance for the ordinal data, we use the Mann-Whitney U test and for the interval data we use the Student's t-test, both with $\alpha = 0.05$.

Attention of the participants. A third user-related factor for the (non-)adoption of recommendations is the lack of attention of the participants. Figure 8 shows how many participants used the recommendations (67 %) and summarizes how the participants answered in case they did not use them. Only 3 % stated that they did not notice the recommendations at all (denoted as `RecNotNoticed`). Another 10 % stated that they noticed the recommendations but did not pay attention to them (`RecNoAttention`). The remaining 20 % looked at the recommendations but did not use any of them (`RecSeenNotUsed`).

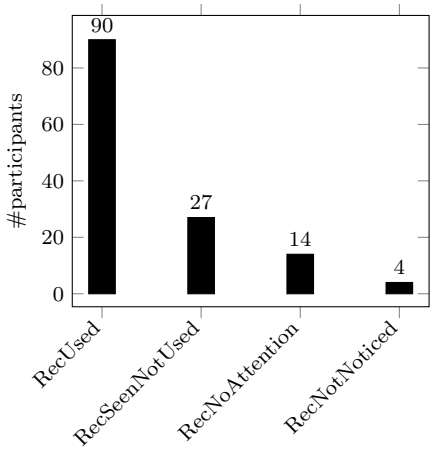


Fig. 8: Distribution of the participants of the `Rec` group with respect to their recommendation use.

With only 3 % of participants from the `Rec` group (4 participants) who did not notice the recommendations, it seems this particular cause of non-adoption is a minor concern. However, the remaining 30 % are more problematic. Possible causes for the non-adoption of the recommendations include (1) the lack of trust in recommender systems, (2) the difficulty of using the recommender and (3) the fact that some participants already knew which tracks to include. These aspects are examined next.

RQ 3: System-related factors for the adoption of recommendations

Trust in recommender systems. One first system-related adoption factor is trust. Based on the answers of the participants to the corresponding question, 25 % of the participants of the `RecSeenNotUsed` group who did not use the recommendations and 7 % of the participants of the `RecNoAttention` group who did not pay attention to the recommendation do not trust recommender systems. Although this number is relatively low, it partially explains why 30 % of the participants who did notice the recommendations chose to not use any of them. Means of increasing trust have been extensively studied in the literature and include the improvement of the presentation (for instance by including

a humanoid agent) and the provision of explanations about the underlying algorithms (Kizilcec, 2016; Berkovsky et al., 2017).

Perceived ease of use. Recommender systems represent one possible solution to reduce the general difficulty of the playlist creation task. However, in order to reduce this difficulty, the effort required to use the service must be outweighed by the benefits of using the service (Armentano et al., 2015).

We asked the participants of both the `NoRec` and `Rec` groups to rate the difficulty of the playlist creation task on a 7-point Likert scale, where 1 means that the participants found it not difficult at all. Both groups assigned a median value of 2. In other words, neither group perceived the task as being difficult, and the recommendations did not increase nor reduce this difficulty, see Figure 9. This observation is in line with our preliminary study, where also no statistically significant difference was found. However, in the previous study the participants of the `NoRec` group found the task slightly more difficult, with a median rating of 3 for the same question (the difference is statistically significant, $p < 0.001$). Compared to the previous study, the definitions of the topics are much more detailed, especially as they make a clear distinction between the corresponding contexts and purposes. Therefore, one explanation for this result could be that defining more precisely in what context and for what purpose a playlist must be created lowers the perceived difficulty of the playlist creation task.

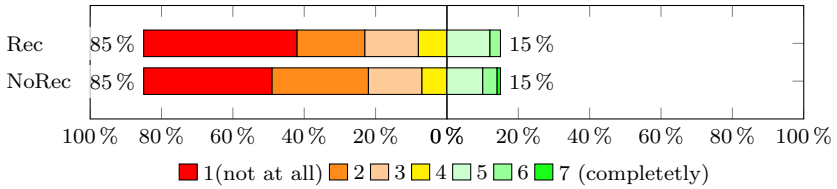


Fig. 9: Distribution of the responses of the participants of the `Rec` and `NoRec` groups to the statement "Overall, it was a difficult task for me to create this playlist."

Overall, the recommendation service did not reduce nor increase the perceived difficulty of the playlist creation task. In other words, the additional effort required to use the service was compensated by its helpfulness, which we study next.

Helpfulness. According to the answers to the questionnaire, 57 % of the participants of the `RecNoAttention` group already knew what tracks they wanted to include, which represents about 6 % of the `Rec` group. In other words, the proportion of participants who did not need any recommendation support was relatively small.

Moreover, 76 % of the participants who used the recommendations believed that recommendation support is *time-saving*, i.e., they responded to the statement "I could save some time using the recommender" with a 5 or higher. This result was confirmed by our objective measures. Looking at the time required to complete the entire playlist creation task shows that the users who used recommendations needed, on average, one and a half minutes less (6.31 minutes) than those without recommendation support (8.03 minutes). The difference is however only marginally significant ($p = 0.07$). At the same time, the participants who used recommendations – and needed less time – created longer playlists, with an average size of 9.38 versus 8.77 for the others.

Overall, the recommendation service was therefore effective in helping the participants find relevant tracks. One specific way of helping them is to provide recommendations of tracks and artists they do not know and that are relevant (Hagen, 2015). We thus asked the participants if the recommendations helped them *discover* such tracks and artists. As shown in Figure 10, 55 % of the participants who used the recommendations stated that the recommender provided relevant tracks they did not know, and 49 % answered that it provided relevant artists they did not know. The service was therefore generally helpful for that purpose. The proportion is, however, around 20 % for the `RecSeenNotUsed` group. This statistically significant difference ($p = 0.001$ for the former and $p = 0.006$ for the latter difference) confirms that *discovery* is an important factor for the adoption of recommendations and therefore confirms the outcomes of previous studies that the importance of this factor in the more general case of music recommendation (Jones, 2010; Lee and Waterman, 2012; Mäntymäki and Islam, 2015).

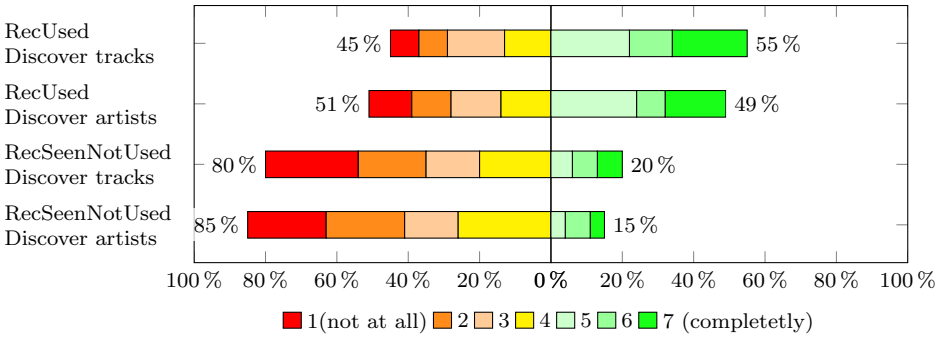


Fig. 10: Distribution of the responses of the participants of the `RecUsed` and `RecSeenNotUsed` groups to the statement "The recommender provided interesting tracks/artists that I did not know."

Criteria	MBC	RF
Homogeneity of musical features, e.g., tempo	449	173
Artist diversity	329	155
Lyrics	258	99
Order	158	68
Popularity	149	56
Transition	147	64
Freshness	121	53

Table 4: Modified Borda Count (MBC) and relevance frequencies (RF): Ranking of playlist quality criteria.

RQ 4: Music-related factors for the adoption of recommendations

We now focus on quality criteria for playlists. We analyze to what extent the participants actually relied on these criteria when adopting recommendations and what other characteristics of the recommendations correlated with the adoption rates.

Quality criteria for playlists. We asked the participants to rank the quality criteria according to their relevance for their playlist (see Section 4 and Table 1 – List 3). Our main goal here was to contrast the expressed statements of the participants with the characteristics of their created playlists and of the adopted recommendations.

In this context, we first analyzed the rankings that were provided by the participants in the post-task questionnaire. To determine the overall ranking, we used the Modified Borda Count method (Emerson, 2013), which can be applied when some rankings are only partial, i.e., when not all items are ranked. We also counted how often each criterion was marked as relevant (relevance frequency). The results are shown in Table 4.

The results indicate that, overall, the participants consider the *homogeneity of musical features* (e.g., tempo, energy or loudness) along with the *artist diversity* of the resulting playlist as the most relevant quality criteria for playlists. Surprisingly, the *lyrics* aspect was ranked third, i.e., before the *order* and the *popularity* of the tracks, as well as *transitions* between the tracks and their *freshness*, which are often considered as major quality criteria in previous studies (Cunningham et al., 2006; Jannach et al., 2014; Dias et al., 2017). One explanation is that the lyrics often correspond to a theme or topic, and that this topic is even more important than the other four criteria. The importance of the topic will be studied in detail later in this section.

Characteristics of the adopted recommendations. The previous result only provides information about what criteria the participants considered most relevant for their playlists. Our next goal is to determine to what extent the participants actually relied on these criteria when adopting recommendations.

However, not all quality criteria could easily be compared with the actual choices of the participants. For instance, our dataset did not contain the lyrics of the tracks, nor the audio signal to assess the quality of the transitions. Still, we could analyze the following criteria. First, more than 45 % of the participants of the `RecUsed` group created playlists in which each artist appeared only once (this value is 57 % across all participants). Although these participants are not the majority, the value can still be considered high as the RAGH recommender always recommended a few tracks from the artists that were already in the playlist. Another criterion we could contrast with the actual behavior of the participants is the order of the tracks. Although the order only came in fourth position in the ranking of the quality criteria, we noticed that about one third of the participants reordered their tracks at least once during the creation of their playlists. This means that this criterion is still quite important, which does not necessarily contradict the ranking of the participants: the order is important for one third of the participants, while the three preceding criteria may be even more important. These two results seem to confirm that the participants actually relied on their provided criteria when adopting recommendations.

Spotify vs. RAGH. The recommendation lists presented to the participants of the `Rec` group contained 10 recommendations obtained via Spotify's API and 10 recommendations that were generated by the RAGH method (see Section 4). Therefore, another way of determining to what extent the participants actually relied on their expressed quality criteria when adopting recommendations is to compare the characteristics of the provided recommendations of each algorithm and the usage of each of these recommendation algorithms.

67 % of the recommendations that were included in the playlists were selected from the RAGH recommendations. This result is in line with the results of the user study reported in (Kamehkhosh and Jannach, 2017) where the recommendations of the CAGH method were often considered to be very suitable for continuing a given playlist. This result confirms the importance of the artists in the selection of tracks, but also suggests that popularity and possibly the user's familiarity with the recommended tracks (as the related artists are more likely to be known) may also be major adoption criteria. However, as our previous results have shown no correlation between the popularity and the adoption rate, it is likely that the major factor that made the RAGH recommendations more acceptable was related to the artists.

Finally, to obtain a better understanding of the characteristics of the provided recommendations of each algorithm, we queried the musical features of the recommended tracks through Spotify's API. Table 5 shows a list of these features. The results show statistically significant differences between the recommendations of Spotify and RAGH in several dimensions, see Table 6. For instance, the recommendations of Spotify are, on average, less popular (avg=33.41) than the recommendation of RAGH (avg=51.07)[17], and are also

[17] The popularity of a track is a value between 0 and 100 (lowest to highest popularity).

Information	**Description**
Acousticness	Absence of electrical modifications in a track.
Danceability	Suitability of a track for dancing, based on various information including the beat strength, tempo, and the stability of the rhythm.
Energy	Intensity released throughout a track, based on various information including the loudness and segment durations.
Instrumentalness	Absence of vocal content in a track.
Liveness	Presence of an audience in the recording.
Loudness	Overall loudness of a track in decibels (dB).
Popularity	Popularity of a track, based on the its total number of plays and the recency of those plays.
Release year	Year of release of a track.
Speechiness	Presence of spoken words in a track.
Tempo	Speed of a track estimated in beats per minute (BPM).
Duration	The duration of the track in milliseconds.
Valence	Musical positiveness conveyed by a track.

Table 5: Description of the collected information for the tracks, as provided by Spotify.

	Spotify		**RAGH**	
Feature	Avg	Std	Avg	Std
Acousticness	0.19	0.29	0.21* ($p < 0.001$)	0.29
Danceability	0.55	0.16	0.55	0.17
Energy	0.71	0.24	0.68* ($p < 0.001$)	0.24
Instrumentalness	0.17	0.31	0.15* ($p < 0.001$)	0.30
Liveness	0.20	0.16	0.19* ($p = 0.001$)	0.16
Loudness (dB)	-7.52	4.34	-7.72* ($p < 0.001$)	4.75
Popularity	33.41	19.38	51.07* ($p < 0.001$)	23.78
Release year	2008	27.69	2006* ($p < 0.001$)	28.26
Speechiness	0.08	0.08	0.08	0.08
Tempo (BPM)	121.76	27.89	122.38	28.79
Valence	0.44	0.24	0.45* ($p = 0.036$)	0.24

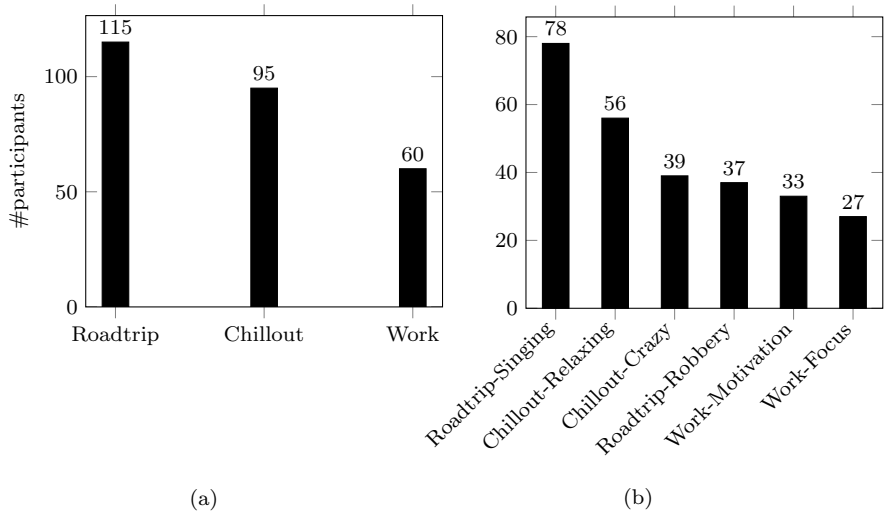
Table 6: Average (Avg) and standard deviation (Std) of the musical features of the provided recommendations of Spotify and RAGH to the `Rec` group (135 participants). The star symbol (*) indicates statistical significance.

released more recently (avg=2008) than those of RAGH (avg=2006), which confirms the low correlation between the freshness and the adoption rate.

RQ 5: Dependency of adoption factors on the topics of the playlist

As explained in Section 4, the participants were presented with six possible topics for their playlists. These topics can be categorized into three different contexts: *chillout*, *roadtrip*, and *work*. For each of the contexts, two different

purposes were considered. As can be seen in Figure 11, some topics were much more popular than others, especially the combination of *roadtrip* and *singing along*, which was selected twice as often as the third most frequent topic.

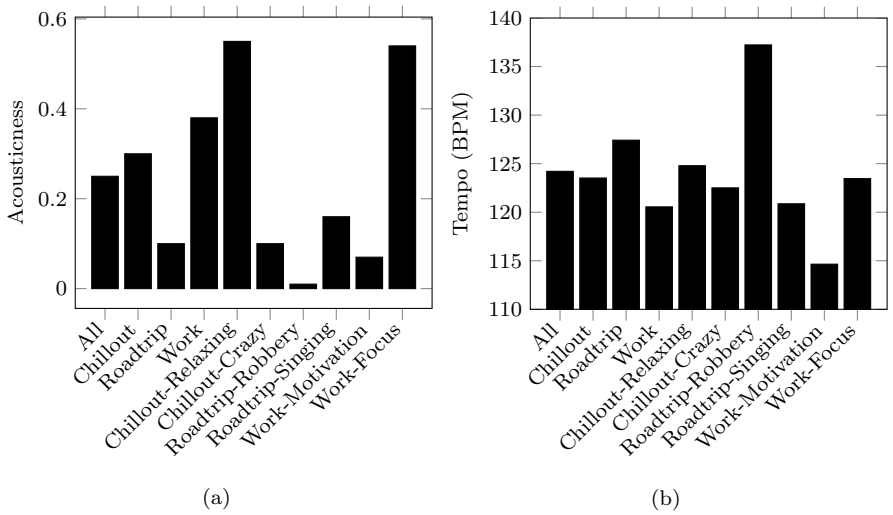


(a) (b)

Fig. 11: Distribution of the selected topics by the participants. The aggregated values for each context are shown in Figure 11a (left), and the detailed values for each topic (context and purpose) in Figure 11b (right).

After the playlist creation task, we asked the participants how well the recommendations matched the topic of their playlist, and computed the correlation between the participants' answers to the question and the number of adopted recommendations. The resulting correlation is 0.18 ($p = 0.046$). Although the decision of using a recommendation was more dependent on the taste of the users (0.35, $p < 0.001$), *matching the topic of the playlist* was still an important adoption factor as the corresponding correlation is comparable to that of artist diversity (0.12, $p = 0.2$) and genre similarity (0.21, $p = 0.02$).

To measure more precisely to what extent the adoption factors depend on the *context* and *purpose* of the playlist, we first compared the respective musical characteristics of the selected recommendations for different topics. The results show differences in various dimensions. For example, on average, the recommendations that were adopted for the *roadtrip* playlists (including both intentions, i.e., *singing along* and *bank robbery*) are significantly less acoustic than the tracks of the playlists for *chillout* and *work* (see Figure 12a). Even in the same context, different purposes can lead to the adoption of recommendations with different characteristics. For example, the average tempo of the selected recommendations for creating playlists with the purpose of *focusing* at *work* is significantly higher than the average tempo of the selected recommendations for *motivation* playlists in the same context (see Figure 12b).



(a) (b)

Fig. 12: Acousticness (a) and tempo (b) of the selected recommendations for different contexts and purposes. According to Spotify, accousticness corresponds to a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence that the track is acoustic.

Another interesting observation in this regard relates to the release year of the selected recommendations. The average release years for the contexts *roadtrip*, *work* and *chillout* are 2002, 2011 and 2012, respectively, i.e., the adopted recommendations for the context *roadtrip* are respectively 9 and 10 years older than the ones selected for both other contexts. Considering the specific purpose of *singing along* in this context, the average is even 4 years older (avg=1998). In other words, users generally do not pick too recent tracks for long journeys on the road, especially when the goal is to sing along. One possible explanation for that phenomenon could be that road trips are often done with a group of friends or with members of the family, in which case the playlist must contain tracks that people of different ages all like. As younger people are more likely to like old songs than older people are to like recent songs (Krumhansl and Zupnick, 2013; Stephens-Davidowitz, 2018), tracks that everybody likes are more likely to be older. The even older tracks for the purpose *singing along* is likely caused by the fact the lyrics must be known: as memorization is dependent on repetition, one is more likely to remember the lyrics of older tracks.

Furthermore, going back to the provided ranking of quality criteria by the participants, but this time with respect to the topics of the playlists, some interesting differences can be observed, see Table 7. For example, the *popularity* aspect was considered a more relevant criterion for *roadtrip* playlists than for *work* or *chillout* playlists. This confirms our previous intuition about

Criteria	All	Chillout	Roadtrip	Work
Homogeneity of musical features, e.g., tempo	449	184	152	113
Artist diversity	329	127	132	70
Lyrics	258	100	115	43
Order	158	65	58	35
Popularity	149	50	76	23
Transition	147	57	49	41
Freshness	121	63	32	26

Table 7: Modified Borda Count: Ranking of playlist quality criteria.

popularity aspects for roadtrip playlists. Similarly, the *transition* aspect was relatively more important for *work* playlists than for the other two contexts.

These differences are in line with the outcomes of previous works showing the importance of the context with respect to the perceived quality of music recommendation (Wang et al., 2012; Schedl and Schnitzer, 2014). The results also show the importance of the purpose, a fact for which, to the best of our knowledge, no evidence existed in the literature.

Overall, in order to better help users during the creation of playlists, service providers should thus probably provide means for their users to indicate the corresponding context and purpose. Furthermore, as the match with the user preferences is currently the most frequent evaluation criterion in academic research of music recommendation, future research could additionally include metrics related to the context and purpose in their evaluations.

RQ 6: Persuasiveness of recommendations

To determine to what extent recommendations influenced the type of tracks the user selected, we first analyzed several characteristics of the tracks depending on whether they were selected from the search interface or from the recommendations. As in previous research questions, we relied on the API of Spotify to retrieve 12 different features for the tracks (see Table 5).

As shown in Table 8, we found statistically significant differences for most of the musical features – exceptions include the liveness, the speechiness, and the tempo feature. The strongest differences we observed were for danceability and popularity ($p < 0.001$): the recommendations picked by the users – across all topics – are less danceable and less popular than what they pick from the search interface. In other words, the adopted tracks were generally quite different from the tracks the participants have selected from the search interface. However, no strong differences were found between the playlists made by participants of the `Rec` group and the playlists made by the participants of the `RecUsed` group. Overall, it seems that the recommendation service merely acted as a facilitator and did not strongly influence the choices of the users of the `RecUsed` group.

Next, we draw our attention to the potential influence of recommendations on the chosen tracks even when users did not use them. We asked the partic-

	Recommended		Searched	
Feature	Avg	Std	Avg	Std
Acousticness	0.25	0.34	0.19* ($p = 0.009$)	0.28
Danceability	0.52	0.17	0.57* ($p < 0.001$)	0.16
Energy	0.68	0.26	0.71* ($p = 0.03$)	0.23
Instrumentalness	0.19	0.33	0.14* ($p = 0.01$)	0.29
Liveness	0.19	0.16	0.19	0.16
Loudness (dB)	-8.20	4.91	-7.45* ($p = 0.01$)	4.42
Popularity	48.79	23.50	54.79* ($p < 0.001$)	18.97
Release year	2008	13.23	2006* ($p = 0.009$)	13.53
Speechiness	0.08	0.08	0.08	0.08
Tempo (BPM)	123.52	29.91	125.70	28.03
Valence	0.42	0.23	0.47* ($p < 0.001$)	0.25

Table 8: Average (Avg) and standard deviation (Std) of the musical features of the selected tracks from the recommendations and the selected tracks from the search interface of the `Rec` group (135 participants). The star symbol (*) indicates statistical significance.

ipants who received recommendations but did not use any of them to state whether the recommendations influenced their track selection. As shown in Figure 13, more than a third of them (37 %) responded positively, i.e., with a rating of 5 or higher. Moreover, we found a strong correlation (0.67, $p < 0.001$) between being influenced by the recommendations and the perceived usefulness of the recommendations. This means that the recommendations were not only useful for the users who used at least one recommendation, but also for a large part of the participants who only looked at them.

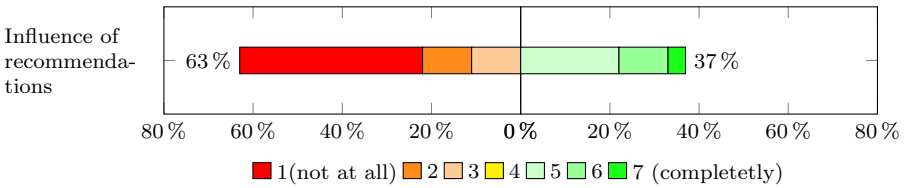


Fig. 13: Distribution of the responses of the participants of the `RecSeenNotUsed` group to the statement "Although I did not use the recommendations, they influenced my selected tracks".

Another indicator for the existence of such an influence is that an average overlap of 74 % could be observed in terms of the artists that appeared in the recommendations that were provided to the participants of the `RecSeenNotUsed` group and the artists of the tracks that they selected manually. This value is 84 % for the `RecUsed` group.

Overall, these results suggest that the mere presence of the recommendations does influence what users select. This effect was previously investigated in user studies in (Köcher et al., 2016, 2018), where the participants exhibited a tendency to select items that were content-wise similar to a (random) recommendation.

6 Summary and Outlook

In this work, we have explored a number of questions related to the adoption and influence of recommender systems during playlist creation. Based on our observations, we derived a number of implications related to the design of such systems which we first discuss in this section. Next, we present the research limitations of our work and future directions.

6.1 Practical implications

Design of User Interfaces. One of our findings is that about 20 % of the participants who did not pay attention to the recommendation or did not use them did not trust recommender systems. Although relatively small, this proportion could be lower by carefully designing the user interface. For instance, Berkovsky et al. (2017) have shown that grouping the recommendations by genre can lead to a substantial increase in trust ratings for movie recommendations. In the same study, explanations about the underlying algorithms were also very efficient, although less efficient than the grouping by genre. These explanations may however have been too much detailed, as Kizilcec (2016) has shown that providing too much information in explanations can reduce their effects on trust improvements.

Our study also sheds light on the importance of the context and purpose of playlists. Although the context has often been in the focus of academic research already, the options provided by commercial services to take the context into account are still limited.[18] Although many platforms allow users to browse music by ambiance, mood, etc., they usually do not let users indicate in what context they intend to play the playlist they are creating, nor with what purpose. It would therefore probably be beneficial for service providers to include such a functionality in their user interfaces.

Design of Algorithms. The homogeneity of track features and the diversity of artists were ranked as the most important criteria for the quality of playlists. Homogeneity is often assumed a particularly important criterion in the literature; however, the criterion of artist diversity is not (Bonnin and Jannach, 2014). Moreover, our analysis of the characteristics of the recommendations

[18] One example of commercial services that provides contextual recommendations and curated recommendations and playlists is the "Listen Now" feature of Google Play Music, see `https://en.wikipedia.org/wiki/Google_Play_Music`.

showed a clear negative correlation between the adoption rate and high levels of artist diversity. This means that recommenders should provide users with recommendations that match the desired artist diversity of the playlist, but which should however not be too diverse.

Another interesting result is that half of the participants who used the recommendation service reported that it helped them discover relevant tracks and artists. This is a strong confirmation that users often wish to discover something new. Still, the other half of participants used the recommendations although they did not discover new relevant tracks. This means that users are generally interested in getting both types of recommendations: (1) tracks that they do not know but may like and that are relevant for the playlists, and (2) tracks with which they are familiar, that they like and that are relevant for the playlist. One implication of that observation is that the recommendations should effectively balance both antagonistic criteria.

6.2 Research Limitations and Threats to Validity

Evaluating the usefulness of recommender systems and analyzing the users' behavior in the presence of such systems with user studies is challenging in different ways. In the music domain, the problem is particularly difficult as it is usually required that the participants listen to a number of tracks during the experiment. This means that they have to invest a considerable amount of effort to complete the study, which may lower the reliability of the results. To alleviate this problem, we provided 30-second previews of the searched and recommended tracks. Since we used excerpts that were selected and provided by Spotify, we are confident that the excerpts are representative of the tracks. Furthermore, we limited the number of minimum tracks in the playlist to six to make the playlist creation task less laborious for the participants. Note that the simplest way of creating a playlist without proper attention is to simply pick one recommendation after the other until the minimum list size is reached. An analysis of the logs however shows that only 13 of the 270 participants could be suspected of a lack of engagement in the study. In the end, as discussed at the beginning of Section 5, the participants were well engaged in the task, which makes us confident that the results are reliable.

Academic user studies in the music domain often have a limited size and in many cases only involve 10 to 20 participants in total (Bonnin and Jannach, 2014). Our study involved 270 participants. The majority of the study participants were university students. While this population of digital natives might be representative of many users of today's digital music services, it remains an open research question to what extent the obtained results generalize to other types of music listeners.[19]

[19] A similar threat to the generalizability of the results relates to the imbalanced gender distribution of the participants. Although we did not ask about the gender of the participants, our approximation by looking at the email addresses shows that the majority of participants are male.

Another limitation in terms of the generalizability of the results is that the acceptance and adoption of recommendations could depend on the quality of the recommender system. We, therefore, built up the user study on two alternative technical approaches. One that is used by one of the market leaders in the music streaming industry[20], and an academic one that has shown to lead to recommendations that were perceived as very helpful by users in previous research (Kamehkhosh and Jannach, 2017). With this selection, we are confident to cover two algorithmic techniques, which are presumably able to generate meaningful and helpful recommendations. One related limitation is that the applied recommender system for the study is not aware of the long-term profile of the participants. This might affect the participants' perception of the usefulness or relevance of the proposed recommendations. We, however, made sure that the provided recommendations were tailored to the selections of the user and were updated every time a new track was added to the playlist.

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

Finally, our study was based on a predefined set of topics and we have to be aware that the choice of the topic might have impacted the observed outcomes. To minimize this threat to validity, we used a systematic procedure – as described above – to identify a set of six topics for which we were confident that they represent no major obstacle for the majority of the participants.

6.3 Future Directions

With our study, our goal is to shed more light on the perception and adoption of recommender systems that are designed to support users in the playlist construction process. Our results showed a high general adoption rate of recommendations during the playlist creation task. Analyzing the responses of the participants to our questionnaire and the characteristics of the adopted recommendations revealed different user-related, system-related and music-related adoption factors. Our results indicated that the relevance of each of these adoption factors depends on the context and intended purpose of the playlist. Finally, our study provided additional evidence for the persuasiveness of recommendations in the music domain.

One of our main observations is that not only the context but also the purpose with which the user creates a playlist has a major importance. In our future work, we plan to study this particular aspect in more detail, especially by proposing algorithms that are able to infer the context and the purpose of a playlist from the available data. One prerequisite is to build a rich dataset in which information about these two particular elements is provided. For

[20] https://musicindustryblog.wordpress.com/2018/09/13/mid-year-2018-streaming-market-shares/

example, the "InCarMusic" data set made available by Baltrunas et al. (2011), despite some limitations (low number of tracks, user ratings of genres instead of individual tracks, one very specific type of context, etc.) can be considered an important step in that direction.

Another interesting finding is the importance of the order. This criterion was ranked comparably high, and one third of the participants reordered their tracks during the creation of their playlists at least once. Current music recommender only recommend tracks to add to the playlist without taking the order into account. One of our perspectives is thus the development of new recommendation models that are able to not only take into account the order of the tracks as input data, but also to suggest positions where to insert the recommended tracks as well as reorderings of the tracks of the playlist.

Finally, our results have shown that recommender systems can also be useful even when the recommendations are not directly adopted, as users can mentally associate a recommended track or artist to a relevant track they know but would not have thought of if they did not have seen the recommendation. Although the usual paradigm is to provide users with items they may like, one research direction that may be particularly relevant is therefore to provide users with items that act as cognitive triggers (Arnott, 2006).

References

Adomavicius G, Bockstedt J, Curley S, Zhang J (2011) Recommender Systems, Consumer Preferences, and Anchoring Effects. In: RecSys '11 – Workshop on Human Decision Making in Recommender Systems, pp 35–42

Andjelkovic I, Parra D, O'Donovan J (2016) Moodplay: Interactive Mood-based Music Discovery and Recommendation. In: UMAP '16, pp 275–279

Andjelkovic I, Parra D, O'Donovan J (2018) Moodplay: Interactive Music Recommendation Based on Artists' Mood Similarity. International Journal of Human-Computer Studies

Armentano MG, Christensen I, Schiaffino S (2015) Applying the Technology Acceptance Model to Evaluation of Recommender Systems. Polibits (51):73–79

Arnott D (2006) Cognitive Biases and Decision Support Systems Development: a Design Science Approach. Information Systems Journal 16(1):55–78

Baltrunas L, Kaminskas M, Ludwig B, Moling O, Ricci F, Aydin A, Lüke KH, Schwaiger R (2011) InCarMusic: Context-Aware Music Recommendations in a Car. In: EC-Web, pp 89–100

Barrington L, Oda R, Lanckriet GRG (2009) Smarter than Genius? Human Evaluation of Music Recommender Systems. In: ISMIR '09, pp 357–362

Baur D, Boring S, Butz A (2010) Rush: Repeated Recommendations on Mobile Devices. In: IUI '10, pp 91–100

Baur D, Hering B, Boring S, Butz A (2011) Who Needs Interaction Anyway: Exploring Mobile Playlist Creation from Manual to Automatic. In: IUI '11, pp 291–294

Berkovsky S, Taib R, Conway D (2017) How to Recommend?: User Trust Factors in Movie Recommender Systems. In: IUI '17, pp 287–300
Bonnin G, Jannach D (2014) Automated Generation of Music Playlists: Survey and Experiments. ACM Computing Surveys 47(2):26:1–26:35
Bostandjiev S, O'Donovan J, Höllerer T (2012) TasteWeights: A Visual Interactive Hybrid Recommender System. In: RecSys '12, New York, NY, USA, pp 35–42
Castagnos S, Brun A, Boyer A (2013) When Diversity Is Needed... But Not Expected! In: IMMM '13, pp 44–50
Chen L, Pu P (2010) Eye-Tracking Study of User Behavior in Recommender Interfaces. In: User Modeling, Adaptation, and Personalization, pp 375–380
Corder GW, Foreman DI (2014) Nonparametric Statistics: A Step-By-Step Approach. John Wiley & Sons
Cunningham SJ, Bainbridge D, Falconer A (2006) 'More of an Art than a Science': Supporting the Creation of Playlists and Mixes. In: ISMIR '06, pp 240–245
Cunningham SJ, Bainbridge D, McKay D (2007) Finding New Music: A Diary Study of Everyday Encounters with Novel Songs. In: ISMIR '07, pp 83–88
DeNora T (2000) Music in Everyday Life. Music in Everyday Life, Cambridge-Obeikan
Dias R, Gonçalves D, Fonseca MJ (2017) From Manual to Assisted Playlist Creation: a Survey. Multimedia Tools and Applications 76(12):14375–14403
Emerson P (2013) The Original Borda Count and Partial Voting. Social Choice and Welfare 40(2):353–358
Fields B (2011) Contextualize Your Listening: the Playlist as Recommendation Engine. PhD thesis, Department of Computing Goldsmiths, University of London
Fogg B (2003) Persuasive Technology: Using Computers to Change what We Think and Do. Morgan Kaufmann
Friedlander JP (2017) News and Notes on 2017 Mid-Year RIAA Revenue Statistics. Online, URL `https://www.riaa.com/wp-content/uploads/2017/09/RIAA-Mid-Year-2017-News-and-Notes2.pdf`
Gourville JT, Soman D (2005) Overchoice and Assortment Type: When and Why Variety Backfires. Marketing science 24(3):382–395
Gretzel U, Fesenmaier DR (2006) Persuasion in Recommender Systems. International Journal of Electronic Commerce 11(2):81–100
Hagen AN (2015) The Playlist Experience: Personal Playlists in Music Streaming Services. Popular Music and Society 38(5):625–645
Hariri N, Mobasher B, Burke R (2012) Context-Aware Music Recommendation Based on Latent Topic Sequential Patterns. In: RecSys '12, pp 131–138
Iyengar SS, Lepper MR (2000) When Choice is Demotivating: Can One Desire Too Much of a Good Thing? Journal of personality and social psychology 79(6):995–1006
Jannach D, Zanker M, Ge M, Gröning M (2012) Recommender Systems in Computer Science and Information Systems–A Landscape of Research. In: EC-Web '12, pp 76–87

Jannach D, Kamehkhosh I, Bonnin G (2014) Analyzing the Characteristics of Shared Playlists for Music Recommendation. In: RecSys '14 – RSWeb Workhop

Jannach D, Lerche L, Kamehkhosh I (2015) Beyond "Hitting the Hits": Generating Coherent Music Playlist Continuations with the Right Tracks. In: RecSys '15, pp 187–194

Jannach D, Kamehkhosh I, Bonnin G (2016a) Biases in Automated Music Playlist Generation: A Comparison of Next-Track Recommending Techniques. In: UMAP '16, pp 281–285

Jannach D, Resnick P, Tuzhilin A, Zanker M (2016b) Recommender systems - beyond matrix completion. Communications of the ACM 59(11):94–102, DOI 10.1145/2891406

Jin Y, Cardoso B, Verbert K (2017) How Do Different Levels of User Control Affect Cognitive Load and Acceptance of Recommendations? In: RecSys '17 – Workshop on Interfaces and Human Decision Making for Recommender Systems, pp 35–42

Johnson C (2014) Algorithmic Music Discovery at Spotify. Online, URL `https://de.slideshare.net/MrChrisJohnson/algorithmic-music-recommendations-at-spotify`

Johnson C, Newett E (2014) From Idea to Execution: Spotify's Discover Weekly. Online, URL `https://de.slideshare.net/MrChrisJohnson/from-idea-to-execution-spotifys-discover-weekly/12-Insight_users_spending_more_time`

Jones N (2010) User Perceived Qualities and Acceptance of Recommender Systems: The Role of Diversity. PhD thesis, EPFL

Jugovac M, Jannach D (2017) Interacting with Recommenders - Overview and Research Directions. ACM Transactions on Intelligent Interactive Systems (ACM TiiS) 7(3)

Kamalzadeh M, Baur D, Möller T (2012) A Survey on Music Listening and Management Behaviours. In: ISMIR '12, pp 373–378

Kamalzadeh M, Kralj C, Möller T, Sedlmair M (2016) TagFlip: Active Mobile Music Discovery with Social Tags. In: IUI '16, pp 19–30

Kamehkhosh I, Jannach D (2017) User Perception of Next-Track Music Recommendations. In: UMAP '17, pp 113–121

Kamehkhosh I, Jannach D, Bonnin G (2018) How Automated Recommendations Affect the Playlist Creation Behavior of Users. In: IUI '18 – Workshop on Intelligent Music Interfaces for Listening and Creation

Kammerer Y, Gerjets P (2010) How the Interface Design Influences Users' Spontaneous Trustworthiness Evaluations of Web Search Results: Comparing a List and a Grid Interface. In: ETRA '10, pp 299–306

Kizilcec RF (2016) How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In: CHI '16, pp 2390–2395

Kjus Y (2016) Musical Exploration via Streaming Services: The Norwegian Experience. Popular Communication 14(3):127–136

Knijnenburg BP, Willemsen MC (2015) Evaluating Recommender Systems with User Experiments. In: Recommender Systems Handbook, Springer, pp

309–352
Knijnenburg BP, Willemsen MC, Gantner Z, Soncu H, Newell C (2012) Explaining the User Experience of Recommender Systems. User Modeling and User-Adapted Interaction (4-5):441–504
Köcher S, Jannach D, Jugovac M, Holzmüller HH (2016) Investigating Mere-Presence Effects of Recommendations on the Consumer Choice Process. In: RecSys '16 – IntRS Workshop, pp 2–5
Köcher S, Jugovac M, Jannach D, Holzmüller H (2018) New hidden persuaders: An investigation of attribute-level anchoring effects of product recommendations. Journal of Retailing
Konstan JA, Riedl J (2012) Recommender systems: from algorithms to user experience. User Modeling and User-Adapted Interaction 22(1):101–123
Krumhansl CL, Zupnick JA (2013) Cascading Reminiscence Bumps in Popular Music. Psychological Science 24(10):2057–2068
Lamont A, Webb R (2011) Short- and Long-term Musical Preferences: What Makes a Favourite Piece of Music? Psychology of Music 38(2):222–241
Lee JH, Waterman NM (2012) Understanding User Requirements for Music Information Services. In: ISMIR, pp 253–258
Lee JH, Bare B, Meek G (2011) How Similar Is Too Similar?: Exploring Users' Perceptions of Similarity in Playlist Evaluation. In: ISMIR '11, pp 109–114
Lehtiniemi A (2008) Evaluating SuperMusic: Streaming Context-aware Mobile Music Service. In: ACE '08, pp 314–321
Lehtiniemi A, Ojala J (2013) Evaluating MoodPic - a Concept for Collaborative Mood Music Playlist Creation. In: IV '13, pp 86–95
L'Huillier A, Castagnos S, Boyer A (2017) Are Item Attributes a Good Alternative to Context Elicitation in Recommender Systems? In: UMAP '17, pp 371–372
Mäntymäki M, Islam A (2015) Gratifications from using freemium music streaming services: Differences between basic and premium users. In: Proc. CIRC
McFee B, Lanckriet GRG (2012) Hypergraph Models of Playlist Dialects. In: ISMIR '12, pp 343–348
Nielsen (2017) Nielsen Music – 2017 Report Highlights. Online, URL `http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2017-reports/us-music-360-highlights.pdf`
Nilashi M, Jannach D, bin Ibrahim O, Esfahani MD, Ahmadi H (2016) Recommendation quality, transparency, and website quality for trust-building in recommendation agents. Electronic Commerce Research and Applications 19:70–84, DOI 10.1016/j.elerap.2016.09.003
North A, Hargreaves D, J Hargreaves J (2004) Uses of music in everyday life. Music perception 22:41–77, DOI 10.1525/mp.2004.22.1.41
O'Keefe DJ (2002) Persuasion: Theory and research, vol 2. Sage
Pauws S (2002) PATS: Realization and User Evaluation of an Automatic Playlist Generator. In: ISMIR '02, pp 222–230
Pichl M, Zangerle E, Specht G (2015) Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name? In: ICDMW

'15, pp 1360–1365
Pontello LF, Holanda PHF, Guilherme B, Cardoso JaPV, Goussevskaia O, Silva APCD (2017) Mixtape: Using Real-Time User Feedback to Navigate Large Media Collections. ACM Trans Multimedia Comput Commun Appl 13(4):50:1–50:22
Pu P, Chen L, Hu R (2011) A User-Centric Evaluation Framework for Recommender Systems. In: RecSys '11, pp 157–164
Quadrana M, Cremonesi P, Jannach D (2018) Sequence-Aware Recommender Systems. ACM Computing Surveys
Schedl M, Schnitzer D (2014) Location-aware Music Artist Recommendation. In: MMM '14, Springer, pp 205–213
Schedl M, Zamani H, Chen CW, Deldjoo Y, Elahi M (2018) Current Challenges and Visions in Music Recommender Systems Research. International Journal of Multimedia Information Retrieval 7(2):95–116
Scheibehenne B, Greifeneder R, Todd PM (2010) Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. Journal of Consumer Research 37(3):409–425
Shiffrin RM, Nosofsky RM (1994) Seven plus or minus two: A commentary on capacity limitations. Psychological Review 101(2):357–361
Sinha RR, Swearingen K (2001) Comparing Recommendations Made by Online Systems and Friends. In: DELOS '01
Steck H, van Zwol R, Johnson C (2015) Interactive Recommender Systems with Netflix and Spotify. Online, URL `http://de.slideshare.net/MrChrisJohnson/interactive-recommender-systems-with-netflix-and-spotify`
Stephens-Davidowitz S (2018) The Songs That Bind. New York Times URL `https://www.nytimes.com/2018/02/10/opinion/sunday/favorite-songs.html`
Stumpf S, Muscroft S (2011) When Users Generate Music Playlists: When Words Leave Off, Music Begins? In: ICME '11, pp 1–6
Swearingen K, Sinha R (2002) Interaction Design for Recommender Systems. In: DIS '02
Tintarev N, Lofi C, Liem CC (2017) Sequences of Diverse Song Recommendations: An Exploratory Study in a Commercial System. In: UMAP '17, pp 391–392
Vall A, Dorfer M, Schedl M, Widmer G (2018) A Hybrid Approach to Music Playlist Continuation Based on Playlist-Song Membership. arXiv preprint arXiv:180509557
Wang X, Rosenblum D, Wang Y (2012) Context-aware Mobile Music Recommendation for Daily Activities. In: MM '12, pp 99–108
Xiao B, Benbasat I (2007) E-commerce product recommendation agents: Use, characteristics, and impact. MIS Q 31(1):137–209
Yakura H, Nakano T, Goto M (2018) FocusMusicRecommender: A System for Recommending Music to Listen to While Working. In: IUI '18, pp 7–17
Yoo KH, Gretzel U, Zanker M (2012) Persuasive Recommender Systems: Conceptual Background and Implications. Springer Science & Business Media.

Zanker M, Bricman M, Gordea S, Jannach D, Jessenitschnig M (2006) Persuasive Online-Selling in Quality and Taste Domains. In: EC-Web '06, pp 51–60

Zhang J, Liu D (2017) Visual Analyses of Music History: A User-Centric Approach. CoRR abs/1703.07534