Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders

Ungruh, Robin; Dinnissen, Karlijn; Volk, Anja; Pera, Maria Soledad; Hauptmann, Hanna

DOI
10.1145/3640457.3688102

Publication date
2024

Document Version
Final published version

Published in
RecSys '24

Citation (APA)
Ungruh, R., Dinnissen, K., Volk, A., Pera, M. S., & Hauptmann, H. (2024). Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In T. di Noia, P. Lops, T. Joachims, K. Verbert, P. Castells, Z. Dong, & B. London (Eds.), *RecSys '24: Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 169-178). ACM. https://doi.org/10.1145/3640457.3688102

Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders

Robin Ungruh
TU Delft
Delft, The Netherlands
R.Ungruh@tudelft.nl

Karlijn Dinnissen
Utrecht University
Utrecht, The Netherlands
k.dinnissen@uu.nl

Anja Volk
Utrecht University
Utrecht, The Netherlands
a.volk@uu.nl

Maria Soledad Pera
TU Delft
Delft, The Netherlands
M.S.Pera@tudelft.nl

Hanna Hauptmann
Utrecht University
Utrecht, The Netherlands
h.j.hauptmann@uu.nl

ABSTRACT

Popularity bias is a prominent phenomenon in recommender systems (RS), especially in the music domain. Although popularity bias mitigation techniques are known to enhance the fairness of RS while maintaining their high performance, there is a lack of understanding regarding users' actual perception of the suggested music. To address this gap, we conducted a user study (n=40) exploring user satisfaction and perception of personalized music recommendations generated by algorithms that explicitly mitigate popularity bias. Specifically, we investigate item-centered and user-centered bias mitigation techniques, aiming to ensure fairness for artists or users, respectively. Results show that neither mitigation technique harms the users' satisfaction with the recommendation lists despite promoting underrepresented items. However, the item-centered mitigation technique impacts user perception; by promoting less popular items, it reduces users' familiarity with the items. Lower familiarity evokes discovery—the feeling that the recommendations enrich the user's taste. We demonstrate that this can ultimately lead to higher satisfaction, highlighting the potential of less-popular recommendations to improve the user experience.

CCS CONCEPTS

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

KEYWORDS

Recommender Systems, Music, Popularity Bias, Bias Mitigation, Fairness, User-Centric Evaluation

ACM Reference Format:
Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. 2024. Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3640457.3688102

RecSys '24, October 14–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688102

1 INTRODUCTION

Music recommender systems (MRS) are essential for today's music consumption, with streaming platforms contributing 67% to the global recorded music revenue [23]. MRS impact users' decision-making process by helping them navigate the vast array of items based on their consumption behavior [46, 47]. Research has shed light on potential biases inherent in MRS, prompting discussions about fairness and equitable allocation of benefits and resources for all stakeholders affected by MRS [13, 15]. One of the most discussed issues in MRS is popularity bias [10, 12, 24, 26, 44].

Popularity bias describes a phenomenon where already popular items get prioritized over lesser-known content [11]. This bias arises because popular items receive more interactions, leading to more frequent recommendations, further solidifying their popularity. Popularity prejudice typically results in the overexposure of some highly popular items while a majority of lesser-known items remain unnoticed [1, 11, 51], leading to detrimental effects for users and artists. For users, item popularity can be a confounding factor in the algorithm's prediction of the user's interest [53]. Popularity bias might cause suboptimal recommendations, especially for niche users, and decrease desirable properties like diversity [24]. For artists, algorithms may disproportionately promote popular items, further amplifying the difference between known and lesser-known artists. Consequently, less popular artists might unfairly have less chance of being recommended to potential listeners [14].

Several **mitigation techniques** have been proposed [e.g., 3, 4, 7, 52] to minimize the negative effects of popularity bias. These techniques typically manipulate (e.g., re-rank) initially biased recommendation lists to achieve fairer recommendations for both users and artists [7]. Algorithmic assessments often assume positive effects of popularity bias mitigation on user satisfaction [7, 44], but such evaluations often overlook nuanced user experiences [24, 33]. Thus, a crucial gap remains in understanding the true impact on user satisfaction, specifically in comprehending how users engage with and perceive such playlists. This knowledge gap stems from the predominant reliance on offline algorithmic evaluations when measuring the impact of popularity bias mitigation [7, 28], which might overlook the nuanced aspects of satisfaction in user's interactions with MRS.

Therefore, our study addresses two research questions (RQs):

(1) To what extent do different popularity bias mitigation algorithms affect user-centered and item-centered fairness within MRS?
(2) How does popularity bias mitigation in MRS influence users' perceptions and satisfaction of recommendations?

To gain insights into the perceptions of 'real' users, we conduct a user study in a lean-in exploration setting [46], comprehensively exploring the effectiveness of popularity bias mitigation techniques in promoting fairness and their impact on users. We quantify how different popularity bias mitigation techniques influence the user-centered and item-centered fairness of an MRS using collaborative filtering, a prevalent recommendation approach [30, 43] known for susceptibility to popularity bias [1, 46]. Subsequently, we investigate whether applying these mitigation techniques to popularity-biased recommendations affects users' perception and evaluation of recommendations. Our contribution lies in extending algorithmic studies [7, 28] by integrating popularity bias mitigation techniques into an MRS and providing an in-depth analysis of their impact on user perception. We capture complex underlying factors such as familiarity and discovery perception, which shape the overall user experience [16, 26, 44] and their relationship with perceived quality and satisfaction. Our unique approach involves retrieving user profiles from the Spotify API [48] and generating personalized recommendations using a pre-trained baseline algorithm on a widely used music-focused dataset [45]. Our comprehensive tool for evaluating recommendation algorithms and mitigation techniques within personalized MRS is openly available and adaptable to various other algorithms[1].

2 RELATED WORK

Here, we review related literature on popularity bias, its effects, mitigation techniques, and the perception of users that forms the basis for our study design and aids in interpreting our results.

Popularity Bias and Fairness. In data used to train RS, it is common to find uneven popularity levels [11, 40, 54], i.e., a long-tail distribution [10]. This distribution distinguishes items into head, mid, and tail categories [7, 34, 36]. In recent works, when categorizing items as such, head items are most popular and constitute the top 20% of interactions, tail items are the least popular receiving 20% of the interactions, and the rest are mid items. In practice, the head consists of a few very popular items, while the tail comprises a large portion of the data with minimal user interactions.

Popularity bias leads to a rich-get-richer effect in which popular entities get disproportionally more attention due to their widespread recognition and exposure in everyday settings. Popular items have a higher chance of being recommended, receiving more interactions from users, and thus becoming even more popular within the feedback loop of RS [1], but unpopular items remain unnoticed by the majority of users. The increase of exposure of already popular items is called **popularity lift**. For instance, Abdollahpouri [1] shows that popularity bias in a common dataset [21] (3% of the items take up more than 20% of the ratings) is intensified by different RS where those 3% of the items accounted for 60% to 100% of the recommendations. This phenomenon is observed for various common RS [1, 12, 24, 36, 51] and MRS [20, 32].

Popularity bias may lead to unfairness for both item providers, i.e., artists, and users. Artists experience unfairness from the disproportionate exposure of their content within recommendation systems despite being potentially just as relevant to the users. Specifically, it amplifies the visibility and accessibility of already well-known artists and their work. As a result, lesser-known or emerging artists face significant challenges in gaining exposure and recognition for their content [14]. Consumers of MRS might be negatively affected since traditional algorithms will produce less optimal results for users interested in less popular items than for other users focused on mainstream items [2]. For instance, a user listening to mainly unpopular items would receive recommendations with a higher proportion of popular items [5]. In contrast, a user who mainly listens to popular items might receive more fitting recommendations in terms of popularity. This creates unfairness regarding the calibration of popularity for certain users [6, 49]. A simple categorization divides users into *mainstream* users being the 20% of users with the highest ratio of head items in their listening history, *niche* users being the 20% of users with the lowest ratio of head items, and *diverse* users being the rest of the users. Studies find that niche users receive recommendations deviating more severely from their listening history [5, 32], which is associated with less fitting recommendations for them [31].

Popularity Bias Mitigation. To address the unfair aspects of popularity bias [15], recent work has introduced mitigation techniques that focus on either user-centered mitigation, such as Calibrated Popularity (CP) [7], or item-centered bias mitigation, such as `FA*IR` [52], aiming to improve fairness metrics for their respective stakeholder group [7, 28]. While item-centered techniques increase the exposure of tail and/or mid items to create a more balanced distribution between under-represented and over-represented items, user-centered ones minimize miscalibration by matching the distribution of popular and unpopular items in the recommendations to the distribution in the user's listening history [7]. Although the goal of fairness methods is to achieve fairness for their respective group, these methods achieve improvements for the other group as well [7]. Algorithmic evaluations have revealed positive effects on fairness when mitigating popularity bias in a user- and item-centered manner [5, 6, 28, 32], but most mitigation approaches are expected to decrease accuracy [28] and utility [31].

Perception of Popularity in MRS. Applying mitigation techniques to RS manipulates the structure of the recommendation lists, typically by re-ranking the initial biased lists created by a traditional recommendation algorithm. Klimashevskaia et al. [27] show that applying a mitigation technique in a movie recommender does not necessarily reduce the relevance of recommendations in a real-life setting. However, the manipulation of a list's popularity inevitably influences the perception and experience of users, as demonstrated in previous studies in MRS [16, 17, 19, 35]. In those, users seem to be able to perceive algorithmic differences in item popularity [35]. Graus and Ferwerda [19] noted that users can perceive the difference between a playlist of highly popular items and a playlist of unpopular items. Some studies have linked differences in perceived

[1] https://github.com/rUngruh/mitigatingPopularityBiasInMRS

popularity with decreased satisfaction [19], but other works have failed to replicate this effect [17, 35].

Such differences in impact may be user- or context-dependent. Ferwerda et al. [16] found that users liked music recommendations if they were familiar with them or when they would foster discovery. In this context, discovery describes to which extent participants feel that the playlists allow them to discover new music and whether the recommendations allow them to refine, depend, and broaden their musical taste [16, 19]. Graus and Ferwerda [19] offers insights into familiarity as a mediating factor responsible for the perceived popularity impacting experience for users with certain personal characteristics. They show that perceived popularity is associated with familiarity in two different ways. Familiarity can positively impact user satisfaction; yet, users with high musical engagement associate familiarity negatively with discovery; an effect that can negatively impact satisfaction.

Recommendation Settings. The setting or use case of the listening session [46] may impact whether discovery will be evoked. Users might appreciate familiar songs during *basic music recommendations* generated to engage users in a listening session with easy navigation, or during *lean-back listening* characterized by users listening to musical recommendations in the background without direct interaction. In contrast, in a *lean-in exploration* setting, the user is motivated to interact with the system to explore music for immediate or future listening. The latter is often referred to when describing the creation of personal music collections or playlists and requires more control and user interaction [46]. Consequentially, low-popularity recommendations achieved through mitigation techniques could lead to higher satisfaction in lean-in exploration sessions. Prior user studies on MRS have not set a specific setting for the user. Some ask users to rate displayed recommendations on a track and a playlist-based level without listening [17, 35]. Others enable exploring the songs by the recommended artists [16] or free listening to previews of the recommended lists [19].

3 METHOD

To investigate the effects of popularity bias-mitigated recommendations on user satisfaction and perception, we conducted a user study. There, we probed participants' perceptions of personalized recommendations that were manipulated by either a user-centered or item-centered mitigation technique. Our study aims to scrutinize the potential of fairer recommendations in lean-in exploration settings. The study was conducted in compliance with ethical and regulatory guidelines associated with human subjects research at Utrecht University, for which an ethical review[2] was conducted. To support reproducibility and follow-up studies on other configurations, we make this work's code openly accessible.

3.1 Setup

We first describe the components that are necessary for exploring the effect of the different recommendation strategies. A visualization of the pipeline that is applied for completing the generation of the recommendations can be seen in Figure 1.

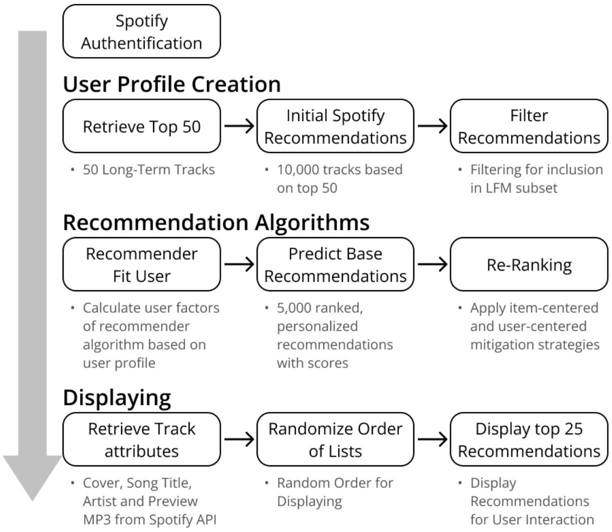

Spotify Authentification
User Profile Creation
Retrieve Top 50
Initial Spotify Recommendations
Filter Recommendations
• 50 Long-Term Tracks
• 10,000 tracks based on top 50
• Filtering for inclusion in LFM subset
Recommendation Algorithms
Recommender Fit User
Predict Base Recommendations
Re-Ranking
• Calculate user factors of recommender algorithm based on user profile
• 5,000 ranked, personalized recommendations with scores
• Apply item-centered and user-centered mitigation strategies
Displaying
Retrieve Track attributes
Randomize Order of Lists
Display top 25 Recommendations
• Cover, Song Title, Artist and Preview MP3 from Spotify API
• Random Order for Displaying
• Display Recommendations for User Interaction

Figure 1: The pipeline used for creating personalized recommendations based on participants' Spotify listening history.

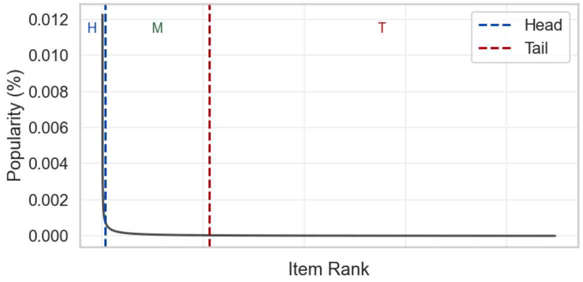


Figure 2: Distribution of user-artist interactions as the ratio of all interactions (in %) by item, ranked by their number of interactions, on LFM-2b$_{filtered}$. 76.44% of the items are tail items (T), 22.99% mid items (M), and 0.57% head items (H).

Data. We use the LFM-2b dataset [45] as a foundation for our analysis and to create user profiles and recommendations. It provides listening events for user-song pairs from February 2005 until March 2020. In line with the procedures of the algorithmic analyses of popularity bias presented in [36, 38], we exclude user-track interactions with a play count ≤ 1 and items and users with less than 5 interactions from the original dataset. We also convert play counts into implicit feedback. Additionally, we only retain songs that are annotated with a matching Spotify URI, which serves as a unique identifier for the songs for the Spotify API. Our final dataset, LFM-2b$_{filtered}$, consists of $171,668,326$ user-song interactions, from $117,337$ users and $2,238,656$ songs. This set is much larger than sets in related studies (e.g., [34]). As shown in Figure 2, LFM-2b$_{filtered}$ has a clear popularity bias, with a small fraction of head items receiving a disproportionately large portion of interactions.

User Profile Creation. To identify users' preferences, we generate User Profiles, which are typically comprised of users' past listening

[2] The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences classified this research as low-risk with no fuller assessment required.

events. Their representativeness depends on the number of available listening events. For instance, Ferwerda et al. [17] require at least 2, 000 events for participants of their user study. However, we utilize the Spotify API to extract preferences, which provides only the top 50 songs per user. This could result in a non-representative profile, i.e., one not sufficiently matching a user's taste. To expand the User Profile with similar songs, we retrieve the top 50 tracks from the long-term history and generate recommendations by using the top tracks of the user as seed songs for input into the recommendation function of the Spotify API[3]. Since the Spotify recommendation algorithm is well-tuned and established, we assume that it provides songs that are similar to the seed songs, thus reflecting the participants' interests accordingly. By repeatedly retrieving new recommendations, we compile a list of 10, 000 songs. We then filter out duplicates and exclude items that are not in LFM-2b$_{filtered}$. Using this approach, we gather at least 250 unique songs for each participant, with a majority of lists including more than 1000 items. We assume that the listening events required by Ferwerda et al. [17] included various duplicate user-track pairings, hence, we argue that this number of unique songs is sufficient. As an extension of the users' top tracks, the lists generated this way will be treated as our *User Profiles*.

Recommendation and Mitigation. For each User Profile, we create three **recommendation lists**. The **base recommendations** are created using RankALS [50], a collaborative filtering algorithm that produces ranked recommendations. We chose this algorithm as it was also used as a baseline in related offline studies [7, 28]. For mitigation purposes, we turn to two recommendation strategies, given their proven performance and ability to improve fairness for their respective stakeholder groups [7, 28]. The user-centered strategy **Calibrated Popularity** (**CP**) [7] focuses on improving calibration between users' listening history and their recommendations. It re-ranks RankALS recommendations by considering the individual preferences regarding the ratio of head, mid, and tail items in the user profile. It aims to create a similar distribution of item popularity in the recommendation list as observed in the User Profiles. The item-centered strategy **FA*IR** [52] supports fairness by guaranteeing a certain exposure of 'protected items' in the final recommendations by re-ranking RankALS recommendations. In this work, we refer to tail items as the protected group, in line with [28].

RankALS was trained with 80% (selected randomly) of the interaction data of LFM-2b$_{filtered}$, using the other 20% as the test set. We selected a factor size of 128 and 30 training iterations. For each generated User Profile, the top 25 predictions of RankALS serve as base recommendations. For the recommendation lists that apply mitigation techniques, we use RankALS to generate a candidate list of 5000 items, as pre-tests have shown that this size ensures the inclusion of necessary tail items for effective re-ranking. Afterward, we apply the re-ranking of the candidate list in line with the mitigation technique's goal; the re-ranked list consists of 25 items. To emphasize the aims of the mitigation techniques in terms of promoting calibration or highlighting underrepresented (tail) items and observe more noticeable differences between recommendations, we choose high strengths of mitigation (CP: $\lambda = 0.99$; FA*IR: $p = 0.98$).

3.2 Procedure

For our study, we followed Knijnenburg et al.'s [29] approach of conducting user studies with an experimental task to probe the user experience in RS. Via mailing lists, we recruited participants with a Spotify account that they used for at least four months. Participants were sent a link to the online study environment, starting with a consent form followed by the Spotify login. During the User Profile and Recommendation creation, users also filled in an initial survey, inquiring about demographic information (age, gender), as well as their musical engagement and sophistication based on Goldsmiths Musical Sophistication Index [39].

After the survey, participants were asked to complete three playlist listening sessions in a tool developed for this study (see Figure 3). In each session, they were presented with the same experimental task of selecting 5 songs from one of the recommendation lists consisting of 25 items. They were asked to choose items they would hypothetically like to add to their playlists for personal listening. This task was chosen so the user interaction with the playlists during the study would resemble an exploration setting, in which the user inspects a playlist to select candidates for immediate or future listening. The three lists with the different recommendation strategies (base recommendation, FA*IR, or CP) were presented in a randomized order. Users could freely browse the provided ranked list by listening to previews of the songs, provided by the Spotify API. If retrieving the preview was not possible, the full song was provided instead. Metadata (song title, artist, and cover) was displayed as well. Users could only complete the listening session after selecting exactly 5 songs. After each session, participants were given a survey. The included questionnaires aim to understand participants' perceptions, attitudes, and evaluations of the recommendations. The questions are asked on a 7-point agreement scale, where 1 indicates "completely disagree" and 7 indicates "completely agree". Upon study completion, users' access to the study tool was revoked, and any personal data used for tool access was deleted.

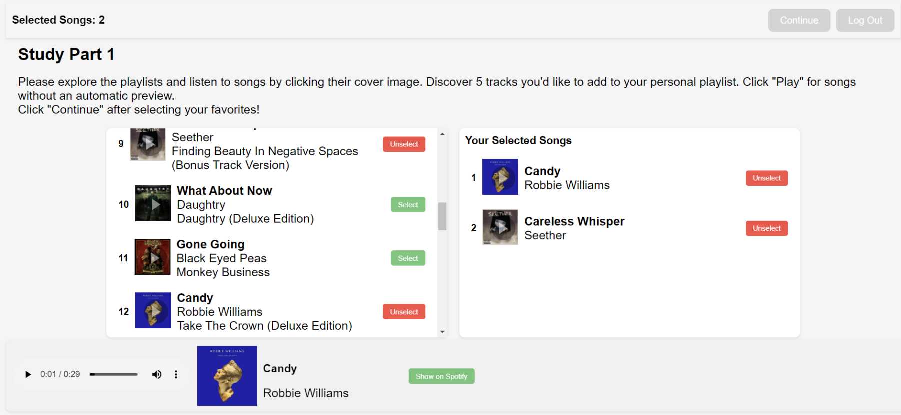
During this study, besides collecting demographic information and questionnaire responses after each session (via surveys), we also retrieved 7 item lists as previously described in Section 3.1: i) one list of > 250 items representing the user profile (*User Profile*); ii) 25 recommendations for each of the three sessions (*Recommendation*); and iii) the 5 items chosen based on the experimental task during each session (*Choice*). Those lists provide insights into the participants' characteristics (like their interest in popular items), the objective system aspects, which explain the structure of the recommendation lists, and the choices of the users, which explain which songs are preferred by the user.

3.3 Measures

The measures utilized to assess the user experience and to gain insights into the structure of the recommendations represent different concepts of the **user's experience** [29, 42]. These concepts are derived by aggregating responses to various questions.

To link the **Objective Systems Aspects** (e.g., song popularity) to the user experience, we measure **Subjective System Aspects**, which represent users' perception of the recommendations. They reflect whether users perceive differences between recommendation lists at all and are typically expected to mediate the effect of

[3] https://developer.spotify.com/documentation/web-api/reference/get-recommendations



Selected Songs: 2

Study Part 1

Please explore the playlists and listen to songs by clicking their cover image. Discover 5 tracks you'd like to add to your personal playlist. Click "Play" for songs without an automatic preview.
Click "Continue" after selecting your favorites!

9 Seether
Finding Beauty In Negative Spaces (Bonus Track Version) Unselect

10 What About Now
Daughtry
Daughtry (Deluxe Edition) Select

11 Gone Going
Black Eyed Peas
Monkey Business Select

12 Candy
Robbie Williams
Take The Crown (Deluxe Edition) Unselect

Your Selected Songs

1 Candy
Robbie Williams Unselect

2 Careless Whisper
Seether Unselect

0:01 / 0:29

Candy
Robbie Williams

Figure 3: The interface for interactions with the recommendations during the study. On the top, the number of selected songs is displayed, with a summary of study instructions below. The center contains two columns: ranked recommendations on the left, and selected songs on the right. On the bottom, a player UI element shows song previews being played.

the objective aspects (like mean popularity or tail ratio) on the **User Experience** [29]. We use questionnaires from Graus and Ferwerda [19] to measure the concepts *Perceived Popularity* (how popular users consider the recommended songs to be), *Familiarity* (whether users know and listen to the presented songs), and *Discovery* (to what extent the songs in the playlist enable the participant to discover new music, to broaden and deepen their musical taste). *Perceived Recommendation Quality* inquires about the users’ perception of the RS’ ability to compute personalized recommendations.

Experience Factors represent the user’s satisfaction and evaluation of the system. Specifically, *Recommendation Satisfaction* [17, 19] asks for general satisfaction with the playlist and its attractiveness, whereas *Choice Satisfaction* [29] measures satisfaction with and enjoyment of songs that the user selected. *Perceived System Effectiveness* [29] assesses how effective and useful the system is.

To assess participants’ interests in engaging with fairer recommendations in the future, we evaluate the constructs of **Behavioral Intentions**. *Choice Listening Intention* (adapted from “purchase intention” [42]) describes the intention to consume the selected items in future listening sessions. *Openness to Similar Recommendations* provides insights into the user’s interest in receiving and listening to items similar to the recommended ones [8].

To gain insights into the structures of the User Profiles, Recommendations, and Choices, we computed various **popularity metrics**. For each of the 7 lists described in Section 3.2 (1 for the User Profile, 3 for the different Recommendations, and 3 for the Choices in each session) the *mean and median popularity* of the selected tracks are retrieved. Popularity is measured as the number of users who interacted with a track in the original dataset. Furthermore, the *ratio of head, mid, and tail items* in the lists is computed based on the categorization of tracks in LFM-2b$_{filtered}$. Finally, for the Choices and Recommendations lists, a comparison to the User Profiles is made, which enables insights into the technique’s impact on popularity. The metric *Popularity Lift* (also called *ΔGAP*) [5, 6, 31] indicates the amplification or reduction of popularity based on the mean popularity of the recommendations, with positive values indicating an increase in popularity. *User Popularity Deviation*, measured by the *Jensen-Shannon Divergence* [18], measures the miscalibration of the recommendation lists [7, 34, 35]. It indicates how dissimilar the (Choice or Recommendations) item lists are compared to the User Profiles, by comparing the ratios of head, mid, and tail items between the two lists. These metrics primarily assess item fairness by examining item exposure across popularity levels, except for the *User Popularity Deviation* metric, which evaluates user fairness.

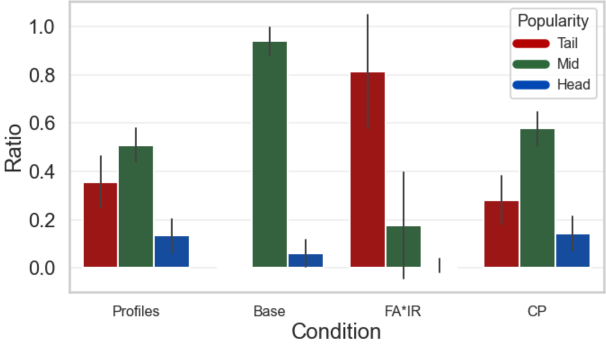


Figure 4: Average distribution of popularity categories in User Profiles and recommendation lists.

4 RESULTS

In this section, we present the outcomes derived from the user study. We make all collected data publicly available[4].

4.1 Participants

Overall, 40 users participated in the experiment. 23 identified as male, 17 as female. We excluded participants who did not complete all surveys and those who did not pass the attention check included in the first post-session survey. The mean participant age was 27.31 ($SD = 9.13$, $min = 22$, $max = 73$). One user did not provide their age.

[4] Data/surveys: https://osf.io/ksqxh/?view_only=1ab775959b784804b2943260e98010dd

Scores for musical sophistication and musical engagement showed a trend towards the center of the scale, with notable variation among participants (MS: $M = 0.085$, $SD = 1.258$; ME: $M = 0.206$, $SD = 0.942$; on a scale from −3 to +3). Based on the classification of users in LFM-2b$_{filtered}$, 21 users can be classified as diverse users and 19 as niche users. No mainstream users were identified.

On average, the User Profiles consist of 1122.20 ($SD = 442.10$) unique tracks per participant. The mean popularity (average of 687.69, $SD = 316.63$) and median popularity (average of 155.86, $SD = 68$) are notably lower than the popularity of User Profiles in LFM-2b$_{filtered}$, which has an average mean popularity of 1246.25 and an average median popularity of 586.77. The lower median popularity compared to the mean suggests generally low popularity among most items, with a few highly popular items that inflate the mean. Popularity ratios indicate an over-representation of tail items and an under-representation of head items in the User Profiles (see Figure 4), with an average ratio of 35.71% tail items, 50.91% mid items, and 13.38% head items, in comparison to the user profiles in LFM-2b$_{filtered}$ (20% tail, 80% mid, and 20% head on average).

All popularity metrics computed on the User Profiles show lower popularity of items in the profiles in our study in comparison to the LFM-2b$_{filtered}$ dataset. Although this deviation indicates a potential limitation of the profile creation approach, we still observe differences in the popularity of recommendations among User Profiles. Hence, we presume that our approach achieves the generation of personalized User Profiles, which are suitable for the creation of personalized recommendations.

4.2 Algorithmic Effects

We first analyze the impact of mitigation techniques on the objective system aspects[5]. Compared to the User Profiles, Recommendation created by RankALS are more popular. The base algorithm does not increase mean popularity (Profiles: $M = 687.69$, $SD = 316.63$, base: $M = 661.47$, $SD = 165.53$) but increases median popularity markedly (Profiles: $M = 155.86$, $SD = 68$, base: $M = 584.08$, $SD = 189.52$). The ratios of popularity categories indicate many mid items ($M = 0.94$, $SD = 0.057$), some head items ($M = 0.06$, $SD = 0.06$), and no tail items in the base Recommendations.

We investigate the impact of mitigation techniques by comparing the popularity of the 3 recommendation lists generated for each user, using the 7 popularity metrics. Multiple ANOVAs demonstrate that these techniques have a significant impact on the popularity metrics of the Recommendations. We analyze those effects further using paired t-tests.

Both FA*IR and CP statistically significantly reduce the mean and median popularity, with FA*IR exhibiting a particularly substantial reduction, with an average mean popularity of 178.71 ($SD = 190.51$) in comparison to 661.47 ($SD = 165.53$) ($t(101) = 17.00$, $p < .001$) in the base recommendations. In contrast, CP only reduces the mean popularity to an average of 604.32 ($SD = 194.66$) ($t(101) = 3.52$, $p < .001$). Similarly CP causes a slight reduction of the median popularity ($t(101) = 4.37$, $p < .001$) and FA*IR leads to a more crucial significant reduction ($t(101) = 14.8$, $p < .001$).

This is reflected in the popularity categories (cf. Figure 4). FA*IR removes the majority of head items ($M = 0.10$, $SD = 0.07$) while CP adds some ($M = 0.14$, $SD = 0.03$). Mid items appear less in the lists re-ranked by both mitigation techniques (FA*IR: $M = 0.18$, $SD = 0.22$; CP: $M = 0.58$, $SD = 0.07$). While CP adds some tail items ($M = 0.28$, $SD = 0.10$), the majority of the items in the playlists created by FA*IR are tail items ($M = 0.81$, $SD = 0.24$). The user-centered method CP promotes head and tail items, which are both underrepresented in the initial recommendations. This is reflected in a low *User-Popularity Deviation* ($M = 0.01$, $SD = 0.01$), indicating a high match between the User Profile and Recommendations. In contrast, the item-centered method FA*IR does not consider the User Profiles and reaches much lower mean and median popularity as well as higher scores in long-tail exposure. For User-Popularity Deviation, only a slight reduction can be seen for FA*IR ($M = 0.22$, $SD = 0.08$) compared to the baseline ($M = 0.25$, $SD = 0.07$).

Multiple regressions show that the popularity metrics computed on the 25 Recommendations significantly ($p < .001$) impacted the distribution of popularity in the final 5 songs selected by the users. For instance, the only significant predictor for *Choice: Mid Ratio* is *Recommendations: Mid Ratio* ($\beta = 0.828$, $t = 7.358$, $p < .001$). The same applies to *Choice: Head Ratio* being predicted by *Recommendations: Head Ratio* ($\beta = 1.163$, $t = 3.969$, $p < .001$).

4.3 User-Centric Evaluation

To explore how techniques influence user perception and satisfaction, we examined participants' views on their interactions with the study tool, based on their survey responses.

CFA. We conducted a confirmatory factor analysis (CFA) using the items' post-session questionnaire scores to validate the constructs created from the questionnaires presented in Section 3.3. After removing items with high cross-loadings and low communalities, we extracted the CFA model, which showed a good fit: $\chi^2(263) = 329.396$, $p < 0.01$, $CFI = 0.993$, $TLI = 0.991$, $RSMEA = 0.046$, 90% CI: $[0.028, 0.061]$. The average extracted variance (AVE) and Cronbach's alpha (α), indicating convergent validity, showed good values ($AVE > 0.5, \alpha > 0.8$), according to typically used cutoff values [22, 29]. The square root of the AVE for each construct should be higher than any of the factor loading of the respective construct to reach discriminant validity. This criterion was fulfilled for each construct except for *Discovery*.

Direct Effects of Popularity on Satisfaction. We conducted ANOVAs to check for any potential direct effects of the condition (i.e., Base, FA*IR, CP) on the three user satisfaction metrics. No statistically significant differences were found (*Recommendation Satisfaction*: $\eta^2 = 0.005$, $F(2, 117) = 0.315$, $p = .731$; *Choice Satisfaction*: $\eta^2 = 0.006$, $F(2, 117) = 0.376$, $p = .687$; *Perceived System Effectiveness*: $\eta^2 = 0.005$, $F(2, 117) = 0.281$, $p = .756$).

Structural Equation Model. By creating a structural equation model (SEM) using the constructs and each condition, according to the guidelines by Knijnenburg et al. [29], we aim to gain more insights into the underlying effects that moderate satisfaction with recommendations. The model (see Figure 5), only including significant effects, showed a good fit: $\chi^2(331) = 425.752$, $p < .001$, $CFI = 0.964$, $TLI = 0.959$, $RMSEA = 0.049$, 90% CI: $[0.034, 0.062]$.

[5] Note that we provide further statistical tests regarding effects between popularity measures in the different lists in the supplementary material.

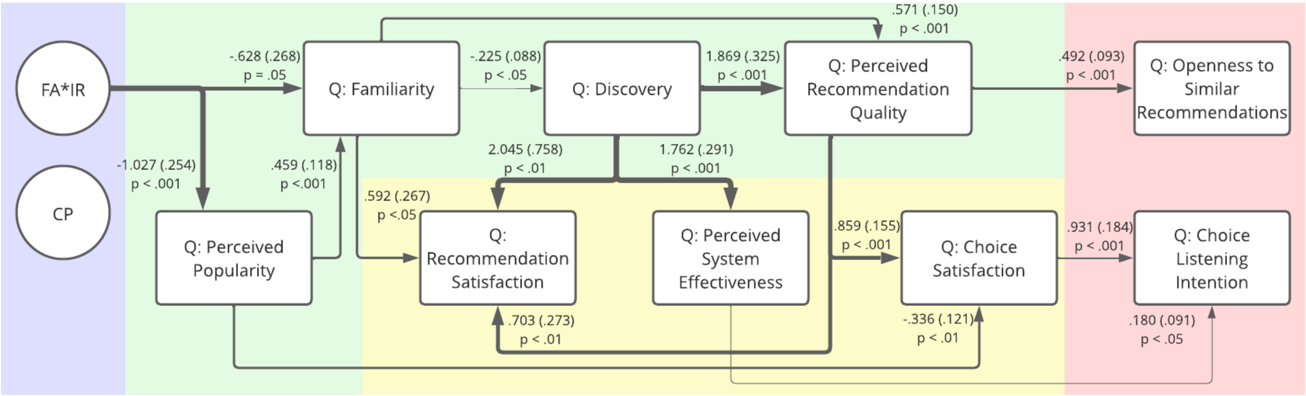


Figure 5: Structural Equation Model constructed from the subjective factors (surveys) and the objective factor, the condition (as indicated by the applied mitigation technique: CP or FA*IR). Numbers next to the arrows describe the standardized coefficients, standard error, and significance.

The model shows that participants' perception of the items recommended by mitigation technique CP does not differ from their perception of the base recommendation items. In contrast, FA*IR has a notable negative effect on *Perceived Popularity* and *Familiarity*. Generally, familiarity has a positive effect on the perceived quality of the recommendations, but familiarity can also reduce discovery, as shown by a negative effect. In other words, low familiarity can impact the perceived quality of the recommendations negatively, but it can also increase discovery, which is another strong indicator for *Perceived Recommendation Quality*. The implications of these findings are further discussed in Section 5.

5 DISCUSSION

We discuss our algorithmic and user-centric results on popularity bias mitigation and user perception with their future implications.

Algorithmic Effects. We investigate how the recommendation algorithms and mitigation techniques impact the popularity of the generated recommendation lists. The popularity metrics on the recommendation lists reveal an expected presence of a popularity bias in the base algorithm. This is especially highlighted by the higher median popularity and the absence of tail items in the base recommendations. This indicates that, compared to the User Profile distribution, many more items of higher popularity were recommended, suggesting that items of low popularity were not seen as valuable recommendations. This is expected from an algorithm affected by popularity bias. Nonetheless, surprisingly, only a few head items were recommended, in contrast to related work in which the same base algorithm, RankALS, tends to recommend mainly head items [7]. This can be attributed to several factors. Firstly, the User Profiles differed in structure from those used to train the model, potentially influencing RankALS' performance. Additionally, its dimensionality was limited due to the large number of items modeled in the vector space. A larger model would likely prioritize head items more heavily. Moreover, the underlying structure of the dataset may impact the RankALS' behavior, as it only spans from February 2005 to March 2020, defining popularity based on listening events during that timeframe. The interests of participants during the study may differ from those reflected in the dataset. Furthermore, Kowald et al. [32] argue that the popularity lift (ΔGAP) metric, and presumably the related mean popularity metric, may be inappropriate for music-related datasets due to their extensive item counts.

We observe that the *mitigation techniques* achieve their respective goals by re-ranking the initial lists. The user-centered approach, CP, avoids miscalibration by aligning the ratio of each popularity category in the Recommendations with the ratios in the User Profiles. Additionally, CP has a slight positive effect on item-centered metrics such as mean popularity and tail ratio. The item-centered approach, FA*IR, enhances item fairness by prioritizing underrepresented items, thereby improving item-centered metrics while only slightly improving user fairness.

User-Centric Evaluation. We analyze the user experience through participants' survey responses and track choices. Concerning the users' perception of the Recommendations, we did not find direct effects of the mitigation techniques on user satisfaction. This aligns with related work observing various levels of popularity [17, 19, 35]. CP aims to align with users' popularity preferences, suggesting that applying user-centered mitigation should improve satisfaction. Hence, the absence of improvement may initially seem like a negative outcome. However, we offer an alternative interpretation of this finding: we have shown that CP manipulates the recommendations of the traditional algorithm, RankALS, by increasing the number of tail items in recommendations. Despite deviating from the recommendations of a traditional algorithm and promoting lesser-known items, CP does not adversely impact users and simultaneously promotes user and item fairness. One comparable work by Klimashevskaia et al. [27] also does not find any harm to the users' engagement with the recommendations when observing the impact on users by applying CP on movie recommendations. There, its application resulted in similar relevance of the items for users in a real-life setting, despite promoting more tail items. Users interacted similarly frequently with item lists that included more tail items. They argue that this challenges presumptions that lower accuracy by applying mitigation techniques would lead to less relevant and engaging recommendations for the users. In our work, FA*IR also does not adversely affect users' evaluations of the RS, despite further promoting underrepresented songs. Consequently, this lends support to proposals advocating for the enhancement of provider fairness.

We find that, according to multiple regressions, the popularity metrics of the Choices are predicted by those computed on the Recommendations of the respective session. This suggests that users align their Choices with the overall trend of popularity reflected in the Recommendations and underscores that while users may not explicitly prefer playlists with varying degrees of popularity, they still lean towards items that follow the prevailing trend, opting for popular items in popular lists and less popular items in lists with lower overall popularity.

Building on the **SEM** analysis, we now delve deeper into the observed effects and their implications. Although previous work did not find clear indications of whether users could perceive differences in popularity [17, 19, 35], our study finds that participants indeed perceive the changes in terms of popularity when applying item-centered bias mitigation, suggesting that the application of FA*IR provides enough changes to be perceptible. In contrast, CP did not impact *Perceived Popularity*, nor any other factor. This could be because the lower degree of reduced popularity may not be sufficient to impact perception. CP's re-ranking of our initial recommendations, mainly comprised of mid items, may have contributed to the absence of this effect, as the re-ranking added both head and tail items to the list. Therefore, the average perceived popularity of the list created by CP might not change noticeably. The fact that no effect of CP on the users' perception and satisfaction could be found — despite aiming to enhance the user experience by accounting for preferences — could be attributed to the similarity between the re-ranked recommendation lists, with a Jaccard similarity of $M = .470$ ($SD = .090$) between the base and CP, compared to $M = .128$ ($SD = .217$) between the base and FA*IR. However, due to its focus on individuals, CP might still lead to improvements for subgroups, such as users with a niche taste, that were not captured in this study.

We find that both FA*IR and *Perceived Popularity* impact *Familiarity*, with the latter effect aligning to results by Graus and Ferwerda [19]. This shows that users are less familiar with the mitigated list, and with playlists that appear less popular. Intuitively, the less popular items, added by the technique, are lesser known by the user, explaining lower ratings for *Familiarity*.

As described in earlier work [16, 19], *Familiarity* is an indicator of satisfaction or list attractiveness. We observe a similar effect of *Familiarity* on *Perceived Recommendation Quality* and *Recommendation Satisfaction* in our work. Besides this, *Familiarity* has a significant negative effect on *Discovery*, showing that being familiar with the recommendations prevents the user from discovering music and expanding their musical taste. *Discovery* is a strong predictor for *Perceived Recommendation Quality*, *Recommendation Satisfaction*, and *Perceived System Effectiveness*, highlighting its importance. From this, we conclude similar results as previously discussed works [16, 19]. *Perceived Quality* and *Satisfaction* can be increased by two contrary effects: i) by creating high familiarity with the lists (e.g., in non-mitigated recommendations), or ii) if *Discovery* is evoked by reducing *Familiarity* (e.g., by applying item-centered popularity bias mitigation). Graus and Ferwerda [19] conclude from those findings that the most satisfying recommendations can be created by achieving a balance between familiar items and those that enable people to discover their musical taste. Our study suggests that *Discovery* is the most salient predictor for perceiving the recommendations as satisfying, effective, and of high quality. This may be due to the 'lean-in exploration' setting of the experimental task. Particularly *Perceived System Effectiveness* is solely significantly impacted by *Discovery*, highlighting the importance of the latter construct. Since the user's task is to explore music, the system is only effective if the user perceives that they can discover music. Effects of other factors that might impact the importance of discovery, like the user's musical engagement [19], were not captured in this study.

Alongside *Discovery*, *Perceived Recommendation Quality*, which highlights users' perception of how well recommendations match their profile, is another crucial indicator for satisfaction. High-quality recommendations lead to greater overall satisfaction, and easier playlist item selection resulting in higher *Choice Satisfaction*. Interestingly, *Choice Satisfaction* is negatively affected by *Perceived Popularity*, suggesting that including several items that seem popular prevents the user from choosing satisfactory items. This could be attributed to the user's goal to find new music, making popular songs seem less attractive since they are already known.

Users express positive behavioral intentions if they are satisfied with the recommendations. This is, for example, indicated by the effect that if recommendations are of high quality (high *Perceived Recommendation Quality*), users desire similar recommendations in terms of genre, style, and artists. These tracks may also reflect similar quality, resulting in higher *Openness to Similar Recommendations*. The *Choice Listening Intention* is also positively influenced by the perceived effectiveness of the system.

Limitations and Future Research. We acknowledge the limitations of our strategy for creating User Profiles, as those in our approach deviate from the actual listening histories of profiles in the LFM-2b dataset. Despite not using an organic listening history, we expect our approach to create user representations suitable for RS use, as i) user profiles rarely reflect users' natural listening histories and are often impacted by feedback loops, thus including recommended items; ii) we found clear differences between participants regarding their interest in popular and unpopular items, indicating that personal differences were captured; and iii) participants reported in open-ended comments that their recommendations generally matched the genres they liked, indicating that the approach could reflect the users' preferences. Nonetheless, we remark that the user profiles–and, in turn, the recommendation algorithms and mitigation techniques–are directly impacted by Spotify's recommendations and, thus, by potential biases and strategies derived from the recommendation generation process. Our findings regarding the users' perceptions, however, still apply. We document the changes in terms of popularity by recommendation, highlighting how popularity bias manifests despite algorithm behavior deviating from our expectations. Furthermore, we show the effects of mitigation techniques affect those recommendation lists. Resulting changes in popularity are perceived by the participants and we analyzed their effects on their satisfaction. Future studies should aim to improve profile creation to align more closely with training data or explore alternative platforms like last.fm [9], which offers comprehensive listening histories. Furthermore, using a dataset limited to songs until 2020 restricts available data and may thus exclude preferences for more recent songs. The time during which the study was conducted (August 2023 to January 2024) reveals potential deviations in user preferences compared to the songs in the dataset. These deviations might involve preferences for more recent songs and changes in songs' popularity between their popularity as indicated by the number of interactions in the dataset and their popularity at the time users interact with it [25].

In Section 4.1, we discuss our potentially biased participant pool. Instead of having 20% mainstream users as expected per definition [5], we identify no mainstream users, but more niche users than anticipated. This raises questions about the generalizability of our findings. However, we argue that the results and implications are valid considering our user group, which shows a general tendency towards medium musical engagement and sophistication. Future studies should further scrutinize how different types of users might perceive popularity bias and its mitigation differently. In our study, such sub-group analysis was not feasible due to the number of participants. Graus and Ferwerda [19] showed the value of such analyses since they found differences in perception of popularity between expert and non-expert users. Similar subgroup analyses could provide further insights into the prospects of recommendation strategies for certain groups. Additionally, convenience sampling was used for participant recruitment, which might not represent the general audience well. Further insights into which factors of the Recommendations were perceived positively could be obtained by collecting ratings for individual items instead of the entire list and evaluating those.

Exploring more recommendation algorithms and mitigation techniques (e.g., [7, 28]) offers the potential for more comprehensive analyses. For example, employing a recommendation algorithm that prioritizes head items (e.g., [7]) could result in re-ranked lists perceived differently by users. Especially CP may generate more significant changes in the lists. Particularly, investigating new state-of-the-art methods that enhance discovery by mitigating popularity bias [44] could yield valuable insights into facilitating discovery through fairer recommendations.

Finally, our results show that implementing item fairness-related methods could positively impact users' behavioral intentions toward fairer recommendations. Future studies can build upon these results and investigate the effects of fairer recommendations longitudinally (similar to [37, 41]) in order to provide further insights into the potential of fair recommendations to promote fairer listening behavior.

6 CONCLUSION

In this work, we investigated how item-centered and user-centered popularity bias mitigation techniques manipulate initially biased recommendations, and how users perceive those changes. We measured various subjective factors, like perceived popularity, familiarity, and discovery, as well as various satisfaction and behavioral intention metrics. We thus advanced the understanding of user perception regarding MRS by conducting a user study — a critical aspect that has been notably under-explored in RS research [24, 33]. We also provided a tool for conducting user studies to evaluate various other recommendation algorithms and mitigation techniques.

We find that neither of the mitigation techniques leads to worse experiences for the participants, indicating the viability of reaching fairer recommendations by promoting underrepresented items while maintaining high satisfaction. We argue that particularly the item-centered method, which promotes tail items even stronger than the user-centered method, holds the potential to enhance the user experience in a lean-in setting by promoting fairer consumption of music. We demonstrate how fairer recommendations can improve recommendation effectiveness through enhanced discovery. However, we acknowledge the potential risks of reducing familiarity, which can affect user's perception of satisfaction and quality negatively. Consideration should be given to the user type and task context when applying such methods.

Our work underscores the importance of evaluating recommender systems not only based on objective metrics based on recommended items but also by presenting these items to real users. This approach allows researchers to identify properties that are *actually* perceived by users, which can have crucial implications for the future design of recommender systems. Particularly, further discussion and reflection are needed to overcome potential trade-offs between recommendations that suit individual users while also fulfilling goals such as fairness.

Our findings have practical implications for designers and developers aiming to create fairer and more engaging user experiences with music recommendations. By showing the potential of mitigating popularity bias within MRS, we pave the way for a more equitable distribution of attention and resources on music platforms. This shift towards fairer recommendation empowers lesser-known artists, creators, and content producers who may have previously struggled to gain visibility in a landscape dominated by popularity.

REFERENCES

[1] Himan Abdollahpouri. 2020. *Popularity bias in recommendation: A multi-stakeholder perspective.* Ph. D. Dissertation. University of Colorado at Boulder.
[2] Himan Abdollahpouri and Robin Burke. 2021. Multistakeholder recommender systems. In *Recommender systems handbook.* Springer, 647–677.
[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems.* 42–46.
[4] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555* (2019).
[5] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).
[6] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems.*
[7] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* 119–129.
[8] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. 2013. Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management* 49, 1 (2013), 13–33.
[9] CBS Interactive. 2024. *Last.fm.* https://www.last.fm/
[10] Oscar Celma. 2010. *Music recommendation and discovery: The long tail, long fail, and long play in the digital music space.* Springer Science & Business Media.

[11] Òscar Celma and Pedro Cano. 2008. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition.*

[12] Savvina Daniil, Mirjam Cuper, Cynthia CS Liem, Jacco van Ossenbruggen, and Laura Hollink. 2024. Reproducing popularity bias in recommendation: The effect of evaluation strategies. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–39.

[13] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. Fairness in Recommender Systems: Research Landscape and Future Directions. (2022).

[14] Karlijn Dinnissen and Christine Bauer. 2023. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization.* 238–249.

[15] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in information access systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

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

[17] Bruce Ferwerda, Eveline Ingesson, Michaela Berndl, and Markus Schedl. 2023. I Don't Care How Popular You Are! Investigating Popularity Bias in Music Recommendations from a User's Perspective. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23).* 357–361.

[18] Andrew V Goldberg. 1997. An efficient implementation of a scaling minimum-cost flow algorithm. *Journal of algorithms* 22, 1 (1997), 1–29.

[19] Mark P Graus and Bruce Ferwerda. 2021. The Moderating Effect of Active Engagement on Appreciation of Popularity in Song Recommendations. In *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part I 16.* Springer, 364–374.

[20] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting consumption towards diverse content on music streaming platforms. In *Proceedings of the 14th ACM international conference on web search and data mining.* 238–246.

[21] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[22] Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal* 6, 1 (1999), 1–55.

[23] IFPI. 2023. IFPI Global Music Report 2022. https://globalmusicreport.ifpi.org/

[24] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015).

[25] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2020. A re-visit of the popularity baseline in recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1749–1752.

[26] Iman Kamehkhosh and Dietmar Jannach. 2017. User perception of next-track music recommendations. In *Proceedings of the 25th conference on user modeling, adaptation and personalization.* 113–121.

[27] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Astrid Tessem, and Christoph Trattner. 2023. Evaluating The Effects of Calibrated Popularity Bias Mitigation: A Field Study. In *Proceedings of the 17th ACM Conference on Recommender Systems.* 1084–1089.

[28] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Christoph Trattner, and Lars Skjærven. 2022. Mitigating Popularity Bias in Recommendation: Potential and Limits of Calibration Approaches. In *Advances in Bias and Fairness in Information Retrieval: Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers.* Springer, 82–90.

[29] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22 (2012), 441–504.

[30] Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook* (2021), 91–142.

[31] Dominik Kowald, Gregor Mayr, Markus Schedl, and Elisabeth Lex. 2023. A Study on Accuracy, Miscalibration, and Popularity Bias in Recommendations. *arXiv preprint arXiv:2303.00400* (2023).

[32] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42.* Springer, 35–42.

[33] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I Jordan. 2020. Do offline metrics predict online performance in recommender systems? *arXiv preprint arXiv:2011.07931* (2020).

[34] Oleg Lesota, Stefan Brandl, Matthias Wenzel, Alessandro B Melchiorre, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization. In *CEUR Workshop Proceedings*, Vol. 3268. RWTH Aachen.

[35] Oleg Lesota, Gustavo Escobedo, Yashar Deldjoo, Bruce Ferwerda, Simone Kopeinik, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2023. Computational Versus Perceived Popularity Miscalibration in Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1889–1893.

[36] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing item popularity bias of music recommender systems: are different genders equally affected?. In *Proceedings of the 15th ACM Conference on Recommender Systems.* 601–606.

[37] Yu Liang and Martijn C Willemsen. 2022. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In *Proceedings of the 16th ACM Conference on Recommender Systems.* 3–13.

[38] Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

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

[40] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems.* 11–18.

[41] Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2024. Assessing the impact of music recommendation diversity on listeners: A longitudinal study. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–47.

[42] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems.* 157–164.

[43] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook* (2021), 1–35.

[44] Rebecca Salganik, Fernando Diaz, and Golnoosh Farnadi. 2023. Fairness Through Domain Awareness: Mitigating Popularity Bias For Music Discovery. *arXiv preprint arXiv:2308.14601* (2023).

[45] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2B: a dataset of enriched music listening events for recommender systems research and fairness analysis. In *ACM SIGIR Conference on Human Information Interaction and Retrieval.* 337–341.

[46] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2021. Music recommendation systems: Techniques, use cases, and challenges. In *Recommender Systems Handbook.* Springer, 927–971.

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

[48] Spotify AB. 2024. *Spotify for Developers.* https://developer.spotify.com/

[49] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems.* 154–162.

[50] Gábor Takács and Domonkos Tikk. 2012. Alternating least squares for personalized ranking. In *Proceedings of the sixth ACM conference on Recommender systems.* 83–90.

[51] Emre Yalcin and Alper Bilge. 2022. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management* 59, 6 (2022), 103100.

[52] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 1569–1578.

[53] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 11–20.

[54] Xiangyu Zhao, Zhendong Niu, and Wei Chen. 2013. Opinion-based collaborative filtering to solve popularity bias in recommender systems. In *Database and Expert Systems Applications: 24th International Conference, DEXA 2013, Prague, Czech Republic, August 26-29, 2013. Proceedings, Part II 24.* Springer, 426–433.