User Perception of Fairness-Calibrated Recommendations

Gabrielle Alves
Universidade de São Paulo
São Carlos, Brazil

Dietmar Jannach
University of Klagenfurt
Klagenfurt, Austria

Rodrigo Ferrari de Souza
Universidade de São Paulo
São Carlos, Brazil

Marcelo Garcia Manzato*
Universidade de São Paulo
São Carlos, Brazil

ABSTRACT

The research community has become increasingly aware of possible undesired effects of algorithmic biases in recommender systems. One common bias in such systems is to over-proportionally expose certain items to users, which may ultimately result in a system that is considered unfair to individual stakeholders. From a technical perspective, calibration approaches are commonly adopted in such situations to ensure that the individual user's preferences are better taken into account, thereby also leading to a more balanced exposure of items overall. Given the known limitations of today's predominant offline evaluation approaches, our work aims to contribute to a better understanding of the users' *perception* of the fairness and quality of recommendations when these are served in a calibrated way. Therefore, we conducted an online user study (N=500) in which we exposed the treatment groups with recommendations calibrated for fairness in terms of two different item characteristics. Our results show that calibration can indeed be effective in guiding the users' choices towards the "fairness items" without negatively impacting the overall quality perception of the system. We however also found that calibration did not measurably impact the users' fairness perceptions unless explanatory information is provided by the system. Finally, our study points to challenges when applying calibration approaches in practice in terms of finding appropriate parameters.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender systems, Fairness, User Study

ACM Reference Format:
Gabrielle Alves, Dietmar Jannach, Rodrigo Ferrari de Souza, and Marcelo Garcia Manzato. 2024. User Perception of Fairness-Calibrated Recommendations. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 01–04, 2024, Cagliari, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627043.3659558

*Supported by CNPq, CAPES and FAPESP, process number 2022/07016-9.

UMAP '24, July 01–04, 2024, Cagliari, Italy

ACM ISBN 979-8-4007-0433-8/24/07
https://doi.org/10.1145/3627043.3659558

1 INTRODUCTION

A substantial fraction of the content that is prominently presented to users on today's online services is automatically determined by a recommender system, e.g., on e-commerce sites, news portals, media streaming services or social networks. Commonly, such recommender systems are designed to create value both for consumers, e.g., by reducing information overload, and for service providers, e.g., by increasing sales or customer retention [20]. However, recent research more frequently also points to potential risks and possible harmful effects of automated recommendations [25]. A well-studied problem in the news and media domain, for example, is that recommender systems may lead to effects like filter bubbles and echo chambers [17, 35], where users are increasingly exposed to only a certain part of the available information. During the most recent years, more general questions of *biases* and *fairness* in recommender systems have received increased interest in the research literature, see [8, 11, 14, 15, 49] for related surveys.

A common challenge in research on fairness in recommender systems—and in the field of "fair ML/AI" in general [7, 34]—is that fairness is a multi-faceted societal construct. Correspondingly, many notions and definitions were put forward in the literature, see [36], and various metrics algorithms were proposed to quantify the extent of unfairness and avoid unfair recommendations. Within this multitude of perspectives, one widely adopted view on fairness in recommendation is based on the notion of *unfair exposure* of items to users. Such an unfair exposure can be the result of a certain *bias* of an algorithm to recommend some items unduly more often than others. Popularity bias is a prime example of such a phenomenon, where an algorithm "learns" to recommend already popular content in a self-reinforcing way [8], thereby further diminishing the chances of new or niche content being recommended to users. In some cases, such a popularity bias may merely result in lower-quality recommendations, which are less personalized to individual tastes. However, it can become a fairness-related issue as well, e.g., if such biased recommendations make it almost impossible for new providers to enter an existing market, or when the recommendations systematically favor the content of the current majority group [12, 19].

One prevalent technical method discussed in literature to address such biases is referred to as *calibration* [27, 38, 45]. In such an approach, the goal is to ensure that system recommendations align with the user's historical preferences in terms of the *distribution* of certain item attributes. In an early implementation of this concept, Oh et al. [38] tried to better align with users' popularity tendencies through re-ranking. Essentially, users who have shown interest

in both popular and less popular content should receive recommendations that reflect this diversity, rather than being exclusively directed towards popular items post-calibration. Similarly, in the context of movies, recommendations can be calibrated based on various item characteristics, e.g., to avoid that only high-budget productions by large studios are recommended even though the user's taste profile includes independent films as well.

Regarding research methodology within fair recommender systems, one recent survey [11] highlighted an over-reliance on data-driven offline experiments within the research community. While these experiments offer insights into algorithm biases, they fail to capture how users perceive fairness-optimized recommendations in terms of quality and fairness perception. To address this gap, we conducted an online user study involving 500 participants. The control group received recommendations from a standard fairness-agnostic collaborative filtering algorithm. Conversely, treatment groups received calibrated recommendations aimed at enhancing fairness based on two item characteristics. These calibrated lists replaced certain items from the original list with 'fairness items'. Furthermore, we explored to what extent the provision of fairness-related information may impact the perception of the recommendations.

Overall, our study revealed that the fairness calibration was largely effective in the sense that participants in the treatment groups selected one of the provided fairness items as their favorite choice to an extent that is mostly proportional to the size of the calibration intervention. Equally important, the analysis of a post-task questionnaire showed that the calibration process did *not* negatively impact the users' quality perception of the recommendations. In sum, this indicates the feasibility of increasing fairness without compromising recommendation quality, at least in the specific domain of the study. However, our study also revealed crucial challenges when implementing a calibration approach in practice, particularly in selecting suitable parameters for the calibration process, see also [29]. Moreover, and even more importantly, the calibration intervention did *not* impact the participants' *fairness perception*, i.e., when asked, they on average did not find the calibrated recommendations to be fairer than the non-calibrated ones, unless when additional explanatory information was provided. An analysis of qualitative feedback by the participants furthermore indicated that they interpret the concept of fairness in quite different ways, emphasizing the challenges of fairness-related research on recommender systems.

The paper is organized as follows. After discussing related work in Section 2, we present our research questions and the detailed study design in Section 3. In Section 4, we then present the main results of our study and discuss practical implications and research limitations. The paper ends with an outlook on future works.

2 BACKGROUND AND RELATED WORK

Fairness in Recommender Systems. Questions related to the ethical and responsible use of AI/ML technology have become a pressing societal concern.[1] Fairness is a central topic in this context, which is often tied to questions of algorithmic biases and discrimination. A significant example is automated recidivism prediction systems, potentially unfairly targeting individuals based on ethnicity [6]. Similarly, discrimination can manifest in recommendation scenarios, like job suggestions, where algorithms might unintentionally guide women toward lower-paying positions [11]. Fairness concerns also extend to more common applications like multimedia and shopping recommendations, where algorithms may favor popular mainstream content over niche or underrepresented offerings.

As fairness is a complex societal construct, various fairness notions have been proposed in the literature, including group vs. individual fairness, consumer vs. provider fairness, or process vs. outcome fairness, see [8, 11, 14, 49]. Scholars argue against a singular general definition of fair recommendations due to this complexity [14]. Regardless of the specific fairness notion, it is essential in fair recommender systems work to clarify the normative claims and goals underlying fairness-enhancing algorithms.

A prior investigation [48] studied user perceptions of fairness and led to a significant finding: participants expecting unfavorable outcomes were inclined to view the algorithmic decision-making process as less fair, even when outcomes coincided with predictions. This highlights the complexity of fairness. In our research, we aim to expand upon these insights. Specifically, we explore the effectiveness of calibration to enhance fairness without sacrificing system quality. Our study mirrors real-world scenarios, such as a public broadcasting organization striving for equitable content exposure. We concentrate on film recommendations, targeting fairness through two key item characteristics: general popularity and production budget, aligning with the notion of "item (or: provider) fairness" [14].

Calibration-based Fairness Approaches. Various fairness-enhancing recommendation techniques were recently proposed, see [49]. In [11], these techniques are roughly categorized as pre-processing, in-processing, and post-processing approaches. Pre-processing techniques are data-oriented, and they basically address bias in the data before they are used for learning. In in-processing approaches, fairness aspects are incorporated into the learning process, e.g., in the form of a specific loss function. Finally, post-processing approaches operate by adjusting a given recommendation list—typically an accuracy-optimized one—to increase a given fairness target.

Yun et al. [51], for example, examined pre-processing and in-processing methods, evaluating the impact of pre-processing on the initial data and subsequent bias during training. Wu et al. [50] introduced Multi-FR, a recommendation framework considering fairness throughout system development and training. Zhu et al. [52] employed a combination of approaches, using pre-processing to define the concepts of Ranking-based Statistical Parity (RSP) and Ranking-based Equal Opportunity (REO) to assess recommendation probability distributions. Post-processing involved implementing the debiased personalized ranking model (DPR), evaluated through empirical experiments on public datasets. Paparella et al. [39] proposed the "Population Distance from Utopia" (PDU) method for post-processing, selecting the best solution based on multiple quality criteria after obtaining Pareto-optimal solutions.

In our study, we focus on a post-processing approach, which can be applied on top of any underlying *baseline ranking* method based, e.g., on collaborative filtering. Differently from techniques that target at improving an aggregate metric, e.g., to reduce the

[1] These concerns recently led to governmental initiatives such as the European Commission's Artificial Intelligence Act or the US AI Bill of Rights.

average popularity of the recommended items across users, calibration works on the user-individual level. Specifically, the calibration goal is to match the distribution of item characteristics of a given user profile with the recommendations. In our work, we study the effects of calibrating the recommendations either according to user popularity tendencies[2] or to the movie production budget.

While calibration approaches may have limits in terms of the global effects one can achieve [30], their advantage is that they consider individual user preferences. In terms of fairness, this means that a user who likes both blockbuster and independent movies will receive recommendations of both types after calibration, and not only blockbuster movies that a majority of other users may prefer. Thus, calibration can be seen as a measure that also helps to increase individual *consumer fairness* because the recommendation quality is not negatively impacted by the majority taste.

Evaluation of Fairness. Assessing the fairness of recommendations and gauging the impact of fairness-enhancing algorithms presents inherent challenges, primarily due to the subjective nature of fairness within specific contexts. Recent research [11] indicates that scholars typically abstract from contextual specifics, relying instead on computational metrics to evaluate fairness. These metrics often revolve around the popularity and exposure of recommended items or providers. Calibration approaches commonly utilize distance measures, such as Earth Mover's Distance or Kullback-Leibler divergence [3, 27, 38, 45] to assess fairness effects by comparing item distributions in user profiles and recommendations. In the offline experiments that complement our user study, we also rely on such measures.

However, a significant limitation of offline experiments and these metrics is the uncertainty regarding their correlation with user perceptions of fairness. Additionally, the trade-off between fairness enhancement and prediction accuracy complicates the assessment of user satisfaction with recommendations. Our current study aims to address this issue through a user study, seeking to understand how fairness enhancements impact users' overall perception of recommendation quality.[3]

Previous research involving human evaluation of fairness in recommender systems is limited [11], with many studies focusing on group recommendation problems [24, 42, 46] rather than personalized recommendations. One notable study [47] examined the effectiveness of a fairness ranking algorithm on a hiring platform. The authors observed mixed results, where the algorithm that favored job seekers from underrepresented groups was only effective in certain circumstances. While there is some relation to our work, we note that the study in [47] did not focus on personalized recommendations. Another study explored user perceptions of fair recommendations through interviews [43], highlighting the importance of explanations for understanding fairness. Drawing from these insights, our online user study includes a treatment condition with explanations.

[2]We note that existing works in that direction mainly aim at popularity bias, and not directly on fairness. We recall that reducing popularity bias *can* be fairness-enhancing in case there is some normative claim or societal desideratum to foster the recommendation of less popular items; see also [2].

[3]A proposal for an alternative, multifaceted evaluation framework for fair recommender systems can be found in [16].

While some research compares computational metrics with user perception, few explicitly address fairness aspects. For instance, Lesota et al. [32] assesses the relationship between computational and perceived popularity calibration but does not directly consider fairness. Our work contributes to this area by bridging the gap between computational metrics and user perceptions, particularly regarding fairness in personalized recommendations.

3 STUDY DESIGN AND MATERIALS

3.1 Research Questions

Before providing the details of our study design, we recall the central aims of our research and formulate the following research questions (RQs):

RQ-1: Are fairness-calibrated recommendations effective? Fairness-aware recommendation algorithms often face a fairness-relevance trade-off, where increasing the fairness of the recommendations may mean recommending at least some "fairness items" that do not have the highest predicted relevance values for a given user. As a result, while a system might make fair recommendations, users may not accept these due to their limited relevance. Therefore, we consider a fairness-aware system effective if users accept the fairness items they are exposed to a certain extent.

RQ-2: How does fairness calibration impact recommendation quality and fairness perceptions of users? While RQ-1 focuses on the impact of fairness-calibrated recommendations on user behavior, RQ-2 aims at understanding how users would *perceive* the system's recommendations. Given the described potential fairness-relevance trade-off, fairness-calibrated recommendations may negatively impact the users' perception of the recommendations' quality, e.g., relevance. Furthermore, we aim to investigate if users actually consider the calibrated recommendations and, in consequence, the entire system, to be fair.

RQ-3: Do explanations impact the fairness perception of algorithms? Inspired by the recent work on user perceptions of calibrated recommendations [32], we investigate if an explicit general statement by the system on fairness-related aspects would impact how users perceive the recommendations.

3.2 General User Study Design

We explored the described research questions through an online experiment in the form of a between-subjects user study in the movie domain. In the study, participants interacted with a website that was created for the purpose of this research. The general flow of the experiment was as follows, see also Figure 1.

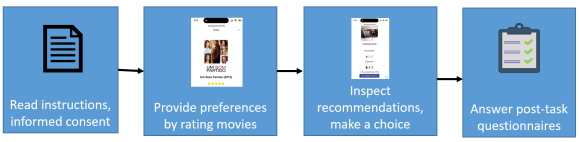

Figure 1: Flow of user study

After reading the instructions and providing informed consent, the user preferences were acquired by asking participants to provide ratings for a number of movies. Based on these preferences, a set of recommendations was shown to the participants, who were asked to select exactly one movie that they would like to watch next. The set of provided recommendations was varied across different participant groups, where the treatment groups were shown different types of fairness-calibrated recommendations, and where the control group received an accuracy-optimized list of recommendations. After making a choice, the participants were asked to answer questions, e.g., regarding their subjective perceptions in a series of questionnaires.

3.3 Preference Elicitation

During this step, participants interacted with an interface with multiple tabs: one was a 'general' tab, showing movies from different genres, and the other tabs showed movies of specific genres. Each tab displayed 12 movies. Furthermore, a search functionality was present. Participants were instructed to rate a minimum of 7 movies from one to five stars. For each movie, the app provided essential information including the title, a poster image, and the year of release.

The movies were selected from the MovieLens-20M rating dataset, which we also used to generate personalized recommendations with the help of a collaborative filtering algorithm in the next step. For our study, we preprocessed the dataset by excluding movies lacking genre information. For the 12 movies featured in each tab, these were kept constant across all users. When gathering preference data, we aimed to minimize popularity bias, thus deliberately eschewing mainstream movies in favor of medium-popularity and niche films. Our selection methodology was balanced: 30% of the movies were randomly selected from the *mid* category, while the remaining 70% came from the less popular *long tail*. Additionally, these movies had an even distribution in terms of budget: one-third were high-budget (*head*), one-third from the *mid*, and one-third were low-budget, *tail* films.[4] To promote impartiality, we randomized the display order of these 12 movies across each tab. This strategy was designed to give every movie, irrespective of its popularity and budget information, an equal chance to be viewed and evaluated by participants.

3.4 Generating Recommendations

We used SLIM [37] as a recommendation model and trained it on the MovieLens dataset. The reasons for this choice are that SLIM not only exhibits very good performance for this dataset [5] but also allows us to make recommendations for new user profiles—as obtained in the preference elicitation step—in a computationally efficient way. As a calibration technique, we used the method proposed by Steck [45], with the Kullback-Leibler divergence as a calibration metric.

Participants were randomly assigned to different treatment groups. The participants in the *control group* in the experiment were shown the accuracy-optimized recommendations as provided by SLIM. The study design included the following five *treatment groups*, which received different types of recommendations:

- **1**-Budget: Recommendations were calibrated to include also movies with a relatively low production budget.
- **2**-Pop: Recommendations were calibrated to also contain less popular movies.
- **3**-Budget+Exp, **4**-Pop+Exp: Same as the previous conditions, but with an explanatory message.
- **5**-Ctrl+Exp: Same recommendations as the control group, but with an explanatory message.

Therefore, while treatments 1 and 2 allow us to gauge the effects of calibration when the system's fairness behavior is a black box to users, treatments 3 to 5 are designed to study RQ-3 on the effects of explanations. Treatment 5 shall, in particular, help us disentangle the effects of explanations and calibration. The explanation messages were shown to users on top of the vertically oriented recommendation lists. For participants in 3-Budget+Exp and 4-Pop+Exp, the text reads[5]: "*Please note that your recommendations may also include [some movies with smaller production budgets | some lesser known movies].*" For the 5-Ctrl+Exp group it read: "*Please note that your recommendations may also include some movies with smaller production budgets and lesser known movies.*"

We emphasize that we deliberately did not include the term "fairness" in the explanations. We believe that explicitly pointing participants to questions of fairness at this stage would influence their fairness perceptions in the end. Instead, we aim to understand if users would connect questions of production budgets and popularity with fairness by themselves.

3.5 User Choice Step

After the preferences were entered, participants were shown a list of 10 recommendations. For each recommended item, we showed the title, release year, and a poster as in the preference elicitation phase. Moreover, participants in all groups were shown some visual indication of the movies' popularity and production budget, as shown in Figure 2. Furthermore, participants could specify if they have already seen this movie, and they could inspect more movie details.

At the end of the list of recommended movies, the participants could select exactly one of the movies as the one that they would like to watch in the future.

3.6 Post-Task Questionnaires

After making their choice, participants were asked to provide answers to a sequence of questionnaires. The first post-task questionnaire was partly based on the validated ResQue framework presented in [40] and partly contained questions that were specific to our study, e.g., regarding the perceived popularity of the recommendations and regarding fairness-related questions. Participants were asked to answer these 14 questions using 7-item Likert-scale responses, ranging from "completely disagree" to "completely agree". Questions **Q1** to **Q14** are listed in Table 1.

[4] Popularity was measured in terms of existing ratings per movie in the dataset. We describe in Section 4.1 how we classified movies into head, mid and tail items based on popularity.

[5] Translated, as the study was not conducted in English.

Figure 2: Fragment of the Recommendation Screen (sketch)

We chose the framework from [40] because it is a well-established and commonly used model in empirical research related to recommender systems [9, 26]. The constructs, including Perceived Quality, Transparency, and Satisfaction, have undergone rigorous validation. This thorough validation process establishes their reliability, making them suitable for inclusion in our study.

After these questions, participants were asked in question **Q15** about what they think would make a movie recommender system fair or unfair. The answer had to be provided in a free-form text field. We note that the question did *not* mention any possible criteria like item popularity or budget aspects.[6] After obtaining the participant's unbiased answer, we asked two specific questions regarding the fairness criteria *popularity* and *production* budget, again using a 7-item Likert-scale item. The questions read:

- **Q16**: I think it is fair when a system also recommends less popular movies occasionally.
- **Q17**: I think it is fair when a system also recommends movies from time to time, which have a lower production budget.

After that, the participants had to complete a final questionnaire about demographics and movie expertise.

4 EXPERIMENT EXECUTION & RESULTS

The experiment was done in two phases. The goal of the first phase was to "calibrate the calibration approach" through offline experiments and a pilot study. In the second phase, the main user study, as described above, was executed.

4.1 Offline Experiments and Pilot Study

As recently reported in [29], applying calibration approaches in practice can be challenging because a number of configuration parameters must be determined for the given application use case. When calibrating for popularity and budget preferences, one central question to answer is when we consider an item to be popular or low-budget. In offline experiments, sometimes threshold values are used based on general heuristics. In [3], for example, the authors apply the Pareto Principle and define 20% most popular movies as the (short) "head", the 20% least popular ones as (long) "tail" items, and the rest to be in the middle. In reality, however, the popularity distribution might be much more skewed, and the authors in [29] for example used a 5% threshold to determine popular items.

Offline Experiments. Prior to the pilot study, we conducted *offline* experiments using the MovieLens dataset to *(a)* select an effective calibration method from the literature and *(b)* determine suitable thresholds to categorize the existing movies into head, mid, and tail items in terms of popularity and budget. Since the MovieLens dataset does not contain budget information, we augmented the dataset with information from IMDb. Movies for which no budget information was found were removed from the dataset, which we share online for reproducibility as well.

Our exploration included recommendation algorithms like SLIM, NeuMF [21], and Mult-VAE [33], and calibration methods such as Steck's method [45], the Personal Popularity Tendency Matching (PPTM) method from [38], and the CP method from [3]. Through evaluating various accuracy, fairness, and popularity measures, we found the combination of SLIM and PPTM to be most effective. However, we later revised our choice of calibration method for the main study, as will be discussed later.

Our refined dataset comprised 7,976 movies, for which we roughly applied the Pareto principle to classify budgets into three distinct categories. The top 18% of movies with the largest budgets were labeled as H (head). The next 57%, with moderate budgets, were classified as M (mid). The final 25%, which included movies with the lowest budgets, were assigned to the T (tail) category. The classification of movies based on popularity was not based on the Pareto principle because of the specific popularity distribution of the data. In this context, a distinct threshold was established, identifying approximately 2% of the movies (totaling 108 films) as head items. These films collectively account for as much as 20% of all user interactions. In contrast, a substantial majority of the films, about 80% (6418 movies), fall into the tail category, yet they only contribute to 20% of the total interactions. The *mid* category encompasses films with moderate popularity, numbering 1448 in total. This group represents a considerable 60% of all interactions.

Using this configuration, we set the thresholds for head, mid, and tail items so that the calibration has a relevant impact on the resulting recommendations. In case calibration only affects a small percentage of the resulting top-10 lists, it is unlikely that we will observe any significant effect in the experiment. Intuitively, our goal was therefore to ensure that, on average, about two items of the top-10 lists, i.e., about 20 percent, shown to users are exchanged with fairness items through the calibration intervention.

Pilot Study Execution and Insights. With these settings, we deployed the pilot study. We recruited 365 study participants by inviting members from a social book review site. We will report more details about the participants' recruitment later when we discuss

[6] The exact question (translated) can be found in the online reproducibility material at https://github.com/user-perception-fairness/umap2024/.

	Question text
Q1	This recommender system gave me good suggestions.
Q2	The system understands my movie taste and preferences.
Q3	Picking just one movie to watch later was hard for me.
Q4	I understood why the items were recommended to me.
Q5	I believe that some movies had a better chance of being recommended than others. *(Fairness)*
Q6	I already know many of the recommended movies.
Q7	I am confident that I have made a good selection.
Q8	There were several good options in the recommendations.
Q9	Please select 'somewhat agree' to show you are paying attention to this question. (Attention check)
Q10	The recommended movies are well-known and popular. *(Popularity perception)*
Q11	The recommended movies were high-budget productions. *(Budget perception)*
Q12	Overall, I am satisfied with this recommender system.
Q13	The recommendations made by the system were generally fair. *(Fairness Perception)*
Q14	I would use a system like this if I needed help finding a new movie to watch in the future.

Table 1: Post-Task Questionnaire Items: Quality and Fairness Perceptions

the main study. We launched the full study, as described in Section 3.2, with a relatively large number of participants in the hope that the configuration determined in the offline experiments based on MovieLens users was, in the best case, appropriate for a real online user population as well.

However, it turned out that this was not the case. When analyzing the data, we observed that the calibration effect in the top-10 lists of online participants was much lower than in the offline experiments. In all treatment groups with calibration, less than 10% of the items were exchanged, and for many participants no item was exchanged at all. We also observed that there were barely any differences between the treatment and control groups, e.g., in terms of the questionnaire responses. Overall, given the small effect of calibration on the recommendations, we could not be sure if the observed non-effect on participants was genuine or due to the small intervention effect.

Our data analysis revealed notable disparities between the user preference models from our study and the typical profiles of MovieLens users. MovieLens users predominantly favor blockbuster movies, whereas our online participants displayed a broader spectrum of preferences, including an appreciation for movies across different levels of popularity. This range can be attributed to our methodology, which introduced users to a wider array of movies, particularly those from less popular categories, during the preference determination phase. Although users had the option to search for specific titles, there was a pronounced tendency to choose films from the middle and tail-end categories. The recommendation algorithm we utilized, SLIM, reflected this breadth in its recommendations, aligning with the varied user tastes we observed. The distinct user profiles in our study likely resulted from our unique approach to eliciting preferences. Additionally, participant behavior might have been influenced by their awareness of participating in a research experiment.

Adaptations. With these insights, we refined our offline experiments to enhance the calibration strategy, aiming for a more impactful intervention. We enriched the MovieLens dataset with user profiles gathered from our pilot online experiment, seeking configurations that would intensify the calibration effect. Our goal was to ensure that roughly 20% of recommended items were substituted with "fairness items" on average. In this process, we focused our parameter search on the online users' data, as their behavior was indicative of the broader audience we aimed to serve in our forthcoming main study.

After exploring and integrating various calibration techniques along with diverse calibration metrics [1, 10, 45], we found out that Steck's method [45], particularly with a constant weight (λ) of 0.5, yielded the desired level of calibration. Both [28] and [45] applied constant trade-off weights between similarity and fairness in their models. We adopted a similar strategy, setting our constant weights λ across a spectrum from 0.0 (pure similarity) to 1.0 (pure fairness), enabling us to manually fine-tune the balance.

As a result of insights gained from these post-pilot offline experiments, we chose to substitute the PPTM calibration method with Steck's method for our main study. Our analysis indicated that adjusting the trade-off weight towards fairness (≥ 0.1) would lead to the replacement of about 3 movies in a list of 10 using this new calibration approach. Table 2 summarizes the calibration effect in different phases.

4.2 Main Study Results

4.2.1 Participants. The main study took place from March 25, 2023 and April 7, 2023. We recruited 500 participants—about 83 per experiment group—by inviting users of a social book review platform in Brazil (90%) and by inviting local university students (10%). The most frequent age group of the participants was "between 22 and 25". Due to the demographics of the user population of the book review platform, the large majority of the participants, over 95%, was female. Participants were invited either via email or by a post on the social platform. The invitation contained a link to the website that was developed for the study. Details about the user demographics can be found in the online material.

4.2.2 Effects on User Choices. To answer **RQ1** on the effectiveness of providing fairness-calibrated recommendations, we analyzed the choice behavior of the participants.

Study Phase	Pre-pilot offline	Pilot Study	Post-pilot offline	Main Study
Avg. N of Exchanges / Top-10	2/10	0.8/10	3/10	2.1/10

Table 2: Average number of items exchanged from top-10 lists during different study phases.

Intervention check. During the preference elicitation phase, participants on average rated 8.64 movies, leading to a solid foundation for personalization. We then compared the distributions of the head, mid and tail items of the movies the participants had rated in the preference elicitation phase with the distribution of head, mid, and tail items they received as recommendations. A χ^2-test revealed significant differences ($p<0.001$) for the non-calibrated participant groups *control* and 5-Ctrl+Exp, whereas the differences were not significant for the calibrated treatment groups. This indicates that calibration was effective in terms of balancing the distributions. As a result, we found that on average about 21% of the recommended items in the calibration groups were fairness items, i.e., they entered the recommendation lists through the calibration process. This indicates a balanced implementation of our calibration process, where most users receive a fair mix of standard and fairness items, without extreme deviations. This rate was very consistent across the four calibration treatment groups, with percentages ranging from around 20 to 23 percent. These outcomes confirm that determining the thresholds based on the data that was collected in the pilot phase was effective. We recall that we aimed at exchanging about 20% of the items and we would not expect that calibration in practice should lead to entirely different recommendation lists as well.

Analysis of Choices. We recall that participants had to select exactly one movie that they would like to watch. When analyzing the actual choices of the participants, we found that in the calibration treatment groups around 20% of the finally selected items were fairness items, i.e., 20% of the users chose one of the items that were added through calibration. Given that also around 20% of the recommended items were calibrated, this indicates that participants considered the fairness items to be equally relevant than those selected solely based on their predicted relevance.

Table 3 shows the details about the participant's choices of calibrated items. We observe no strong difference between the groups regarding their tendency to select calibrated items. This also indicates that the provision of short explanatory statements (3-Budget+Exp and 4-Pop+Exp) may have not largely impacted the participants' awareness of popularity and budget aspects when they made their choices. In terms of the calibration criterion, we found that calibration for popularity mostly directs participants to tail items. In contrast, when considering the production budget, calibration led participants also to select head and mid items, i.e., more costly productions. This is expected given the rating distributions for head, mid, and tail items which we observed in the preference elicitation phase.

Considering the *position* of the finally selected items in the list, we could find no indication of position bias, where participants would mostly pick items from the top of the list. Instead, the distribution of positions of the selected items was quite balanced, both for the control and treatment groups.

4.2.3 User Perceptions: Quality and Fairness. **RQ2** aims to understand how participants perceived the recommendations in different dimensions. To that purpose, we analyzed the responses to the post-task questionnaires. In total, 430 of the 500 participants filled out the questionnaire and passed an attention check (Q9), leaving us with between 67 and 78 participants per experiment group.

To check for differences between the groups, we applied a Kruskal-Wallis test (α=0.05) for each of the 15 questionnaire items[7] shown in Table 1. Significant differences were observed only in two cases: Q10 on popularity (p=0.04) and Q13 on fairness (p=0.01), see Table 4[8]. For these two questions, we applied a Mann-Whitney U post-hoc test, accounting for multiple comparisons, again with α=0.05 as a significance level.

Popularity. For Q10, the post-hoc test revealed a *significant* difference between treatment group 2-Pop and the two non-calibrated groups *Control* and 5-Ctrl+Exp. Specifically, the participants in the popularity-calibrated group indeed perceived the recommendations to be less popular than those of the accuracy-optimized recommendations. On the other hand, all other pairwise post-hoc comparisons were not significant.

Fairness. When asked if the recommendations were generally fair (Q13), the post-hoc test indicated several significant differences, see Table 5. A substantial *effect size* based on *Hedge's g* was however only identified in two cases, notably between *Control* and 5-Ctrl+Exp, and between 1-Budget and 5-Ctrl+Exp.[9] In both cases, 5-Ctrl+Exp is involved, which is the condition where accuracy-optimized lists were presented in the control group; this time, however, with a statement that the list may contain less popular and low-budget productions. The significantly different fairness perception of *Control* and 5-Ctrl+Exp is notable, as it indicates that *participants actually related the general statement on item popularity and budgets to fairness, and that they (unduly) attributed more fairness to the system that provided the statement.*

Looking at the other substantial effect size between 1-Budget and 5-Ctrl+Exp, we found that the fairness perception of 1-Budget is—together with the control group—the lowest across the groups. We recall that the 1-Budget group received budget-calibrated recommendations, where the calibration generally led to the stronger inclusion and selection of higher-budget productions, see Table 3. This increased the inclusion of such movies without explanation but had no impact on the fairness perception of users. Generally, looking at the responses for Q13 across groups, we found that the provision of explanations generally led to slightly higher responses.

[7] There were 17 questions in total (Q1-Q17), but one was open-ended and one an attention check.

[8] The full results are shown in the online material.

[9] Hedge's g, a variant of Cohen's d, corrects the latter's bias in small sample sizes, as Cohen's d often overestimates effect size [22, 41]. This adjustment makes Hedge's g more accurate, especially for smaller studies. Despite similar interpretations for effect sizes, Cohen's benchmarks—0.2 for small, 0.5 for medium, and 0.8 for large effects—should be contextually applied, as the definition of small, medium, or large effects varies across different fields [13].

Treatment Group	Selection from calibration	H	M	T	Total
1-Budget	21%	5	9	3	17
2-Pop	19%	0	1	15	16
3-Budget+Exp	18%	6	6	3	15
4-Pop+Exp	20%	3	3	11	17

Table 3: Percentages of items selected from calibrated items per treatment group

	Control	1-Budget	2-Pop	3-Budget+Exp	4-Pop+Exp	5-Ctrl+Exp	p-value
Q10	5.94 (1.05)	5.84 (1.30)	5.46 (1.36)	5.65 (1.43)	5.66 (1.35)	5.99 (1.27)	0.04
Q13	5.54 (1.54)	5.53 (1.41)	5.97 (1.02)	5.98 (1.38)	5.76 (1.48)	6.12 (1.16)	0.01

Table 4: Means and std. deviations for questions Q10 (popularity) and Q13 (fairness), p-value according to a Kruskal-Wallis test.

This points to an interesting area of future work on the topics of explanations, trust and fairness perceptions, see also [32]. In our present study, the explanations only partly led to a higher fairness perception (for Q13, but not for Q5), and the explanations also did not significantly impact the perceived transparency of the system (Q4). Nonetheless, our findings at least partly answer research question **RQ3**, indicating that explanations can be a mechanism that impacts the fairness perceptions of a system.

Medium effects (Hedge's g)			
Control	vs.	5-Ctrl+Exp	p<0.01, g=-0.42
1-Budget	vs.	5-Ctrl+Exp	p<0.01, g=-0.45
Moderate effects			
Control	vs.	2-Pop	p=0.03, g=-0.32
1-Budget	vs.	2-Pop	p=0.02, g=-0.35
1-Budget	vs.	3-Budget+Exp	p=0.04, g=-0.31

Table 5: Significant post-hoc comparisons and effect sizes (Q13).

Besides the two mentioned contrasts with a substantial effect size, we identified three more contrasts with significant differences, as listed in Table 5. However, the effect sizes are only moderate accordingly.

Other Quality Factors. As mentioned, we only found significant differences between groups for two out of 15 questions. This in particular means that the calibration approaches did *not* negatively impact the quality perceptions of the participants. For example, the participants did not feel that the recommendations were less accurate than in the control group, even though we on average exchanged 20% of the items with the highest predicted accuracy with (still somewhat relevant) fairness items. This finding suggests that fairness calibration can be effectively applied—at least when limited to a certain extent–without leading to negative effects on the quality perception of the system.

4.2.4 User Notions of Fairness. Our final aim was to investigate what participants considered to be aspects of fairness in recommender systems. We recall that we first asked the participants to discuss this topic in their own words, and we then asked them if they would find it fair if the system occasionally included items with lower popularity or with lower production budgets (Q16 and Q17).

Regarding the questionnaire, we found that participants considered both aspects to be highly related to fairness. The average response to the two questions was about 6.5 on the 7-point Likert scale item for both aspects, with no statistically significant differences between the control and treatment groups. Overall, this confirms the intuition that item popularity and budgets can be relevant factors for end users in terms of fairness.

In our qualitative analysis, we used content analysis to explore users' perceptions of fairness in their responses. Responses were coded to develop a category system, adhering to conventional content analysis where categories emerge directly from the text data [18]. This bottom-up coding helped us identify patterns and themes within the responses [23]. Using this coding approach, we identified various notions of fairness within the responses.

In analyzing user feedback regarding fairness of the recommendation system, a noteworthy observation emerges: while many participants perceived the system's suggestions as fair, few could articulate the reasons underpinning this perception, often lacking depth in their explanations. Furthermore, a substantial number of users confessed their inability to define the characteristics of a 'fair' recommendation system.

Despite this, a common theme resonated among most users: the alignment of recommendations with personal preferences as a marker of fairness. This points to the importance of enhancing the accuracy and personalization of recommendations. This viewpoint was encapsulated for example in user *u-112*'s remark: "*A recommendation feels fair when it considers not just personal taste but also the user's current mood. While I generally prefer action and adventure movies, there are times when I'm more inclined towards a comedy.*" This insight underscores the necessity for a recommendation system that dynamically adapts to fluctuating user preferences and personality.

In contrast, users *u-108* and *u-95* offered a different perspective on fairness. They emphasized the importance of diversity in recommendations, beyond mainstream and popular choices. User *u-108* noted, "*Fairness is compromised if suggestions are limited to films from major studios or those receiving significant hype,*" while user *u-95* added, "*Recommendations are fair when they align with my viewing history, but become unfair if they only feature well-known*

movies, ignoring the wealth of lesser-known but thematically similar films from diverse cultures, like Latin or Turkish cinema." These comments highlight a desire for a recommendation system that not only caters to personal tastes but also broadens exposure to a diverse range of cinematic works.

The issue of users being confined within their 'preference bubble' was explicitly addressed by user *u-353*, who observed, "*The movie choices suggested by Artificial Intelligence can be problematic, as they often reinforce the user's existing preferences. For instance, someone who predominantly watches comedies is likely to receive recommendations that are exclusively comedic, perpetuating a narrow content exposure.*" This concern, however, was echoed by only a minority of users, indicating a potential area for further exploration and user education regarding the implications of algorithm-driven content curation.

A related topic mentioned by several other participants is that a system is *unfair if it recommends low-quality movies* or does not recommend high-quality movies. Given that quality may be a subjective factor, this aspect can relate to the previous personalization aspect. However, it generally raises the question of how a recommender system should deal with items that are considered to be of limited quality by the general community, e.g., ones with low average ratings, i.e., should they be considered at all by a recommender system?

Overall, while popularity and budget aspects do play a role for users, we conclude that most participants focused on individual recommendation accuracy as a key factor regarding what end users would relate to fairness in a recommender.

4.3 Discussion

Implications. Overall, our study showed that calibration can indeed be a valuable means to increase the fairness level of recommendations in terms of up-ranking items, which would otherwise not be exposed to (and noticed by users) as frequently, in a personalized way. Our study showed that the recommended *fairness items* are chosen by participants in a proportional way. Therefore, the fairness intervention can be seen to be effective, as certain items were not only exposed more often but also frequently chosen by the participants.

Moreover, an analysis of the user perceptions revealed that the calibration did *not* negatively impact their recommendation quality and fairness perceptions. An important finding in this context, however, is that the *provision of explanations* can be a valuable means to increase the fairness perception of the system. These observations call for more research in the future on fairness, explanations, transparency and trust in recommender systems.

The investigation of what participants would consider a fair recommendation showed that the aspects that are the focus of our study, popularity and budget, are sometimes connected by the participants with fairness. The most important aspect that emerged from the qualitative analysis of the participants' feedback is that recommendation *accuracy* is the predominant factor determining the system's fairness from the end-user perspective.

The significance of accuracy underscores the effectiveness of calibration approaches, which are capable of maintaining an appropriate accuracy level. In contrast, alternative methods, such as those aimed at merely decreasing popularity bias at an aggregate level, might result in a noticeable decline in the perceived recommendation quality. This observation is supported by the analysis of related user studies in [31] or [44]. On the other hand, calibration can be limited in terms of mitigating global biases. If, for example, the large majority of a user community prefers blockbuster movies, then calibration might not reduce the bias much at an aggregate level, see [30], but may, in theory, even reinforce it.

Our study also showed that applying a calibration technique in a practical setting is not trivial, as also discussed in [29], because appropriate thresholds have to be determined, which cannot be easily derived from offline experiments. Our study emphasized the importance of collecting realistic data from a deployed system in a pilot phase. Furthermore, domain expertise is required to find an appropriate level of calibration. If the intervention is too small, the overall effects may be too limited. If, on the other hand, too many less relevant items are shown to users, e.g., to help them explore alternative content, the quality perception may decrease, see also [4].

Research Limitations. A possible threat to the external validity of our work is that our participants are largely young females. This is due to the fact that the social book review site where we sourced them has predominantly female users. More research is thus needed to verify that our findings can also apply to other demographic participants with a longitudinal analysis. Nonetheless, we believe that our participants are representative to a certain extent, at least for a younger generation of today's online users. Furthermore, we are confident that these participants' responses are reliable, given that they were intrinsically motivated and volunteered to participate in the experiment. The large majority of the participants also passed the attention check that we built into the experiment.

From a methodological perspective, in our current experiment, we largely relied on validated questions from the RESQUE framework for the user-centric evaluation of recommender systems [40], which we augmented with additional questions about fairness-related perceptions. We only used one single question item per such construct so as not to overwhelm participants. An extension and further validation of the questionnaire items related to the perception variables and an analysis of their potential relations with other variables remain a part of our future work.

5 OUTLOOK

Fairness has received increased attention in recent years, both in the areas of recommender systems and machine learning. However, many research works on this topic solely rely on computational experiments to assess the effects of fairness-aware algorithms. With this work, we aimed to study such effects with humans in the loop. Our online user study revealed that calibration-based approaches, if tuned properly, can indeed be effective in guiding users' choices without negative effects on the quality perception of the system. Our work focused on budget and popularity-related considerations of fairness in a particular domain. Our future work includes the investigation of fairness-related questions in other domains. Furthermore, we plan to address the question of the longitudinal effects

of algorithmic fairness interventions and how they would affect the individual stakeholders.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, R. Burke, and Bamshad Mobasher. 2019. The Impact of Popularity Bias on Fairness and Calibration in Recommendation. *ArXiv* abs/1910.05755 (2019).

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 726–731.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021*. ACM, 119–129.

[4] Gabrielle Alves, Dietmar Jannach, Rodrigo Ferrari, Daniela Damian, and Marcelo Garcia Manzato. 2023. Digitally Nudging Users to Explore Off-Profile Recommendations: Here Be Dragons. *User Modeling and User-Adapted Interaction* online first (2023).

[5] Vito Walter Anelli, Alejandro Bellogin, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art. In *30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2022)*.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He†. 2022. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* (2022).

[9] Li Chen and Pearl Pu. 2014. Experiments on user experiences with recommender interfaces. *Behaviour & Information Technology* 33 (2014), 372 – 394.

[10] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Syst. Appl.* 181 (2021), 115112.

[11] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction* online first (2023).

[12] 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, UMAP 2023*. 238–249.

[13] Joseph A. Durlak. 2009. How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology* 34 9 (2009), 917–28.

[14] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16, 1-2 (2022), 1–177.

[15] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 679–707.

[16] Mehdi Elahi, Himan Abdollahpouri, Masoud Mansoury, and Helma Torkamaan. 2021. Beyond Algorithmic Fairness in Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. 41–46.

[17] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, Eivind Fiskerud, Adrian Oesch, Loek Vredenberg, and Christoph Trattner. 2021. Towards Responsible Media Recommendation. *AI and Ethics* 2, 1 (2021), 103–114.

[18] Satu Elo and Helvi Aulikki Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62 1 (2008), 107–15.

[19] Daniel Fleder and Kartik Hosanagar. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* 55, 5 (2009), 697–712.

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

[21] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW '17*. 173–182.

[22] Larry V Hedges. 1981. Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics* 6 (1981), 107 – 128.

[23] H. F. Hsieh and Sarah Elizabeth Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15 (2005), 1277 – 1288.

[24] Nyi Nyi Htun, Elisa Lecluse, and Katrien Verbert. 2021. Perception of Fairness in Group Music Recommender Systems. In *26th International Conference on Intelligent User Interfaces*. 302–306.

[25] Dietmar Jannach and Markus Zanker. 2021. Impact and Value of Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Bracha Shapira, and Lior Rokach (Eds.). Springer US.

[26] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. *Proceedings of the 9th International Conference on Human-Agent Interaction* (2021).

[27] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient Optimization of Multiple Recommendation Quality Factors According to Individual User Tendencies. *Expert Systems With Applications* 81 (2017), 321–331.

[28] Mesut Kaya and Derek G. Bridge. 2019. A comparison of calibrated and intent-aware recommendations. *Proceedings of the 13th ACM Conference on Recommender Systems* (2019).

[29] 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 *17th ACM Conference on Recommender Systems (Late Breaking Results)*.

[30] 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*, Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo (Eds.). 82–90.

[31] Kibeom Lee and Kyogu Lee. 2015. Escaping your comfort zone: A graph-Based recommender system for finding novel recommendations among relevant items. *Expert Systems with Applications* 42, 10 (2015), 4851–4858.

[32] 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 (SIGIR '23)*. 1889–1893.

[33] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *WWW '18*. 689–698.

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (2021).

[35] Lien Michiels, Jorre Vannieuwenhuyze, Jens Leysen, Robin Verachtert, Annelien Smets, and Bart Goethals. 2023. How Should We Measure Filter Bubbles? A Regression Model and Evidence for Online News. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. 640–651.

[36] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170. 3.

[37] Xia Ning and George Karypis. 2011. SLIM: Sparse linear methods for top-n recommender systems. In *Proceedings of ICDM '11*. 497–506.

[38] Jinoh Oh, Sun Park, Hwanjo Yu, Min Song, and Seung-Taek Park. 2011. Novel Recommendation Based on Personal Popularity Tendency. In *ICDM '11*. 507–516.

[39] Vincenzo Paparella, Vito Walter Anelli, Franco Maria Nardini, R. Perego, and T. D. Noia. 2023. Post-hoc Selection of Pareto-Optimal Solutions in Search and Recommendation. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (2023).

[40] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 157–164.

[41] Robert Rosenthal. 1984. Meta-analytic procedures for social research.

[42] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Fairness in Package-to-Group Recommendations. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*. 371–379.

[43] Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[44] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the 2011 ACM Conference on Recommender Systems (RecSys '11)*. Chicago, Illinois, USA, 125–132.

[45] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 154–162.

[46] Maria Stratigi, Haridimos Kondylakis, and Kostas Stefanidis. 2017. Fairness in Group Recommendations in the Health Domain. In *33rd IEEE International Conference on Data Engineering, ICDE 2017*. 1481–1488.

[47] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. *Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring*. 989–999.

[48] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[49] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.* 41, 3 (2023).

[50] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2021. A Multi-Objective Optimization Framework for Multi-Stakeholder Fairness-Aware Recommendation. *ACM Transactions on Information Systems* 41 (2021), 1 – 29.
[51] Ye Yuan, Xin Luo, and Mingsheng Shang. 2018. Effects of preprocessing and training biases in latent factor models for recommender systems. *Neurocomputing* 275 (2018), 2019–2030.
[52] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).