

Break Out of a Pigeonhole: A Unified Framework for Examining Miscalibration, Bias, and Stereotype in Recommender Systems

YONGSU AHN, University of Pittsburgh, United States

YU-RU LIN, University of Pittsburgh, United States

Despite the benefits of personalizing items and information tailored to users' needs, it has been found that recommender systems tend to introduce biases that favor popular items or certain categories of items, and dominant user groups. In this study, we aim to characterize the systematic errors of a recommendation system and how they manifest in various accountability issues, such as stereotypes, biases, and miscalibration. We propose a unified framework that distinguishes the sources of prediction errors into a set of key measures that quantify the various types of system-induced effects, both at the individual and collective levels. Based on our measuring framework, we examine the most widely adopted algorithms in the context of movie recommendation. Our research reveals three important findings: (1) Differences between algorithms: recommendations generated by simpler algorithms tend to be more stereotypical but less biased than those generated by more complex algorithms. (2) Disparate impact on groups and individuals: system-induced biases and stereotypes have a disproportionate effect on atypical users and minority groups (e.g., women and older users). (3) Mitigation opportunity: using structural equation modeling, we identify the interactions between user characteristics (typicality and diversity), system-induced effects, and miscalibration. We further investigate the possibility of mitigating system-induced effects by oversampling underrepresented groups and individuals, which was found to be effective in reducing stereotypes and improving recommendation quality. Our research is the first systematic examination of not only system-induced effects and miscalibration but also the stereotyping issue in recommender systems.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:

Yongsu Ahn and Yu-Ru Lin. 2023. Break Out of a Pigeonhole: A Unified Framework for Examining Miscalibration, Bias, and Stereotype in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2023), 19 pages. <https://doi.org/10.1145/3484509>

1 INTRODUCTION

Recommender systems have been widely adopted in a variety of domains, ranging from online shopping to high-stakes decisions, such as insurance or job applications. Using data and recent advances in AI techniques, these systems can generate “personalized” items and information that are tailored to the needs of individual users. Despite its benefits in reducing their browsing and retrieval efforts, studies have found that recommender systems induce biases, such as popularity bias [1], which favors *popular* items, filter bubbles [36] or echo chambers [17], where recommended

Authors' addresses: Yongsu Ahn, University of Pittsburgh, Pittsburgh, United States, yongsu.ahn@pitt.edu; Yu-Ru Lin, University of Pittsburgh, United States, United States, yurulin@pitt.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

2160-6455/2023/1-ART1 \$15.00

<https://doi.org/10.1145/3484509>

items or users in the system are tailored by a group of users of *similar* kind, resulting in item/group segregation. Recommender systems are also likely to amplify user preference for certain item categories (e.g., action or sci-fi movies) that are *dominantly* preferred by a larger group of users (e.g., men) [27, 41].

Compared with the issue of system-induced bias, the existence of stereotypes perpetuated by recommendation systems is relatively unexplored. “Stereotype” refers to the behavior of over-generalizing user characteristics by regarding individual users as having typical preferences of a predetermined group (e.g., based on demographic or other attributes) [8]. For example, in movie recommendations, a system may recommend to male users a set of items that typical male users will like. Such behavior can lead to discrimination against minority subgroups or individuals with particular interests, restricting users’ choices into typical preferences, which in turn reduces user satisfaction or the quality of recommendations.

In this study, we look at the stereotypes and biases that arise as a result of data underrepresentation within a system. Because underrepresented users – those with atypical preferences or who belong to minority groups – are more likely to be impacted, they are more likely to receive poor-quality recommendations, resulting in discrimination against underrepresented subgroups or individuals. These system-induced behaviors can worsen the miscalibration problem, also known as the mismatch between user preferences and specific item categories [39].

While there is collective evidence pertaining to user characteristics [3, 28], systematic behaviors [1, 5, 12, 40], and miscalibration [39], we find that existing literature lacks a comprehensive understanding of the underlying relationships, posing the following challenges: 1) A lack of a systematic approach to identifying the relationship between users, algorithms, and errors: Existing research has contributed to the identification of various types of system-induced biases; however, there is a lack of systematic methods to examine how these system-induced effects are associated with user characteristics and exacerbate recommendation quality. 2) Insufficient comprehension of the stereotyping issue in recommender systems: Unlike other potential system-induced effects, the problem of stereotyping has not been studied in the literature. Stereotyping recommendations can limit the diversity of individual preferences, especially by influencing users with atypical preferences to receive recommendation items that are typically preferred by the groups to which they belong, thereby failing to calibrate the user preferences.

This study aims to provide a unified framework for quantifying stereotypes, bias, and miscalibration, given the challenges and the need for a better understanding of the relationship between them. Using the proposed framework, we are able to analyze miscalibration at a granular level, examine the interconnectedness of system-induced effects, and gain a deeper insight into those systematic effects on both user groups and individuals. We focus on answering the following research questions: 1) **Sources of miscalibration:** To what extent do two distinct sources of error, bias, and variance, contribute to miscalibration? 2) **System-induced effects:** How does it relate to system-induced effect? and 3) **Disparate impact:** How does it affect groups and individuals disproportionately? Based on the proposed measures, we conduct an observational study on existing recommender systems to observe how systematic effects and stereotyping problems result in negative impacts on groups and individual users. Moreover, we investigate potential mitigation strategies for users negatively affected by the system-induced behavior. The contributions of this study are as follows:

- **A unified framework for understanding miscalibration, bias, and stereotype:** We present a statistical framework for decomposing miscalibration into its two distinct sources, bias and variance, as well as their relationship with stereotype. This enables us to investigate further systematic behaviors that lead to various types of effects and have a disproportionate impact on underrepresented groups and individuals.

- **A formulation of the stereotyping issue in recommender systems:** We discuss the problem of stereotyping in recommender systems, which is the tendency to overgeneralize user preferences based on their group membership. This behavior is associated with two system-induced effects, stereotype and inflated diversity, both of which contribute to miscalibration. This is the first study to examine the risk of stereotyping problem in recommender systems.
- **An analysis of the relationship among user characteristics, system-induced effects, and miscalibration in existing recommender systems:** We investigate the presence of system-induced effects in existing movie recommendation systems, as well as their associations with user characteristics and miscalibration. Based on the proposed measures, we characterize the behavior of recommender systems that tend to stereotype user preferences and discriminate against atypical users and minority groups. Our analysis reveals three key findings with implications for mitigating this issue.
- **An investigation into the problem of data underrepresentation:** To explore the possibility of mitigating the disparate impacts of systematic behaviors over minority groups and individuals, we analyze the effect of alleviating data underrepresentation by oversampling users who are stereotyped and explore how it changes the systematic behavior of recommender systems. We discuss its effect on the amelioration of stereotyping and miscalibration, as well as the remaining challenges.

2 RELATED WORK

2.1 Bias and discrimination in recommender systems

It has been studied [1, 5, 12, 40] that recommender systems are susceptible to biases that favor popular items or users, or get stuck on items and users of the similar kinds. Among them, popularity bias [1, 11] is a widely studied systematic behavior unfavorable towards less popular items in the long tail appearing much less than popular items. These biases can be amplified while in the iterative feedback loop [12], leading to rich-getting-richer effect. Studies have captured its consequential effects, such as filter bubbles [5, 40, 44] where homogeneous items are being over-recommended, or echo chambers [36, 40] where groups are more polarized in their preferences or ideologically segregated. These in turn can degrade the quality of recommended items or prevent users from discovering new items.

These biases also tend to have a disparate impact over demographic groups and individuals, discriminating against minor groups or users by offering a lower quality of recommendations. Recent studies have pointed out that these fairness problems in recommender systems are even more complex due to its multi-sided nature [15] in either or both consumer and provider side. For example, in book recommendations, female authors were found to receive lower exposure to customers than male authors [15]. This exposure bias was found to be related to the size or entropy of the user rating profile in [30]. According to the analysis, females tend to have fewer interactions and specific interests, and as a result, receive lower-quality recommendations. While various biases have been studied, our study attempts to view these system-induced biases from the perspective of underrepresentation problem. Focusing on the user side, we investigate how individuals and groups who are underrepresented in the data have a greater impact of bias and discrimination in recommender systems.

2.2 Stereotyping problem in recommendations

Stereotype is a widely known cognitive phenomenon that refers to people's shared beliefs over specific group characteristics. While stereotype can help an intuitive assessment of groups [8],

it can easily lead to overgeneralization of individuals' characteristics [?] based on their group membership. For example, in the movie preferences, [45] found that people tend to cognitively exaggerate group differences with respect to widely known gender stereotypes (e.g., men like action and sci-fi movies) greater than how groups actually differ. Recent research has explored the possibility that machine learning or recommender systems may also be subject to stereotypes. On the bright side, previous research has attempted to utilize stereotype-based user modeling in several different recommender system domains [4, 7, 16, 25]. On the other hand, some of the existing literature [4, 35] argued that a trained classifier can pick up biases that warp the representational space by pulling outlying examples closer to the prototypical preferences of a dominant group. Recent research [27, 33, 41] on recommender systems has investigated bias disparity, a problem related to the stereotyping issue at the category level. It refers to a disparity or amplification of users' preference ratio on a particular item category. In [27, 33, 41], collaborative filtering algorithms were examined to observe bias disparity especially in the exclusively favored genres by males or females such as action and romance movies. To mitigate these biases, several approaches have been proposed to construct and intervene in the causal pathways of user preferences [43] or train recommender systems with dual representations of user preferences [6, 19]. While these studies shed light on the existence of stereotypes, the stereotyping problem in recommender systems has not been thoroughly defined and examined in terms of its effect and relationship with user characteristics. In this study, we introduce the notion of system-induced stereotype and investigate how it affects underrepresented groups and individuals.

2.3 Miscalibration

Miscalibration in recommender systems refers to the discrepancy between actual and predicted preferences over item categories [39]. Compared to item-level metrics, such as Normalized Discounted Cumulative Gain (nDCG), measuring the accuracy of recommendations in top-ranked items, this category-level measure allows evaluating a system's capability of distributional calibration across various areas of interest (i.e., item categories). In the later work, miscalibration was investigated in terms of its relationship with bias and fairness [2, 3, 14]. In [2], they found that a greater impact of popularity bias can lead to an increase in miscalibration. Such an impact was also found to vary across recommendation algorithms and demographic groups.

On the other hand, little has investigated the deeper understanding of miscalibrations, i.e., how it is associated with different sources of error and system-induced effects. Our study introduces a framework for decomposing miscalibration into bias, variance, and stereotype, which allows one to disambiguate how different systematic effects are combined to form the overall miscalibration and how to mitigate them.

2.4 Diversity

Diversity is one of the most highly valued properties in recommender systems, as it allows users to be presented with a variety of unexpected and diverse items. Measures such as intra-list distance [48], or entropy [13, 46], have been extensively used in studying diversity and its bias effect in recommender systems. Several studies have investigated how diversity can be promoted with shared accessibility of user preference representations [6] or by iteratively updating learning objective for diversity in a supervised [13] or unsupervised manner [46]. In our study, we observe that increasing diversity in a recommender system can result in unexpectedly poor recommendations; we refer to this phenomenon as "inflated diversity." We find that diverse items that are liked by typical users are recommended to atypical users with specific interests. Throughout the study, we investigate how this system-induced effect is related to other measures and how it disproportionately affects groups and individuals.

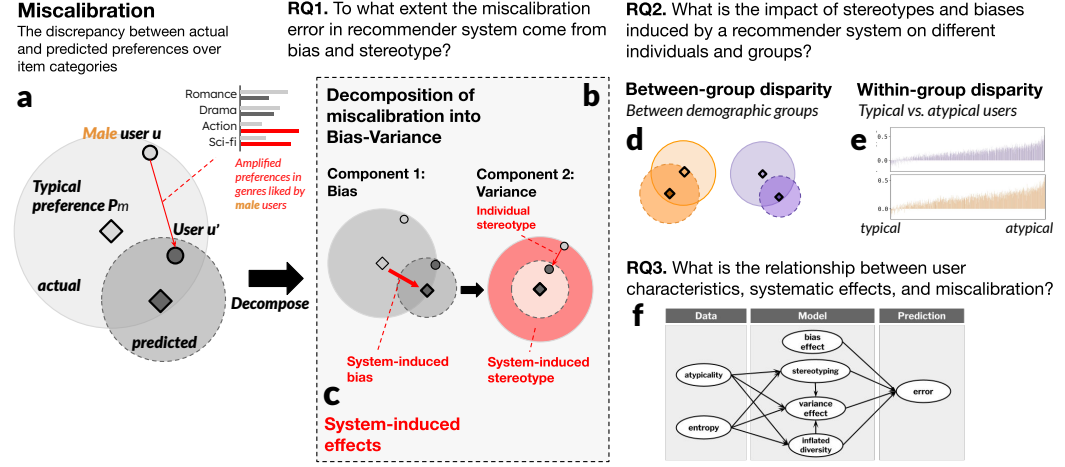


Fig. 1. The overview of the study. We aim to provide a unified framework for the systematically examining miscalibration, system-induced bias and stereotype. (a) Miscalibration is a category-level error in recommender systems, referring to the discrepancy in distributions between the distributions of actual and predicted preferences over item categories. (b) Our framework proposes to decompose miscalibration into two distinct sources of error, bias and variance. (c) It allows us to capture how two system-induced effects, bias and stereotype, can be associated with miscalibration. (d, e) By measuring the group-level or individual-level disparity, we measure whether miscalibration, bias, and stereotype disproportionately impact groups and individuals. (f) The relationship between user characteristics, system-induced effects, and miscalibration can be better understood through association analysis.

3 MISCALIBRATION, BIAS, AND STEREOTYPE

In this section, we present a unified framework for systematic understanding of miscalibration, bias, and stereotype in recommender systems. Fig. 1) provides an overview of our research methodology. As illustrated in Fig. 1a, the process of generating recommendations can be conceptualized as the mapping between the space of users' actual preferences (colored as light gray) and the space of users' predicted preference (colored as dark gray). A miscalibration problem arises when a recommender system fails to predict a user's preference of items at the category level, e.g., a user who does not favor sci-fi movies frequently receives sci-fi movie recommendations. Conceptually, such a mismatch can be understood as a misalignment between two spaces in Fig. 1a. In our framework, this distributional misalignment can be decomposed into two distinct parts of distortions, bias and variance (Fig. 1b). Inspired by the well-known Bias-Variance decomposition in machine learning literature [9, 23], we develop a statistical framework that decomposes miscalibration into bias and variance and further characterizes the system-induced stereotype effect. We show that miscalibration is associated with system-induced bias and stereotype (Fig. 1c). These effects can disproportionately affect groups (Fig. 1d) and individuals (Fig. 1e), resulting in discrimination in terms of recommendation quality.

In the following, we propose a set of measures to allow a deeper understanding of miscalibration. In Section 3.1, we first define the notion of miscalibration error. Next, we show that the miscalibration can be decomposed into its two sources of error using Bias-Variance decomposition under KL-divergence loss (Section 3.2). This allows us to discuss system-induced bias and stereotype as two distinct system-induced effects that are related to bias and variance respectively (Section 3.3).

Notation	Term	Definition	Type
p, q	Actual/predicted preference for a user u	Probabilistic distribution of a user u 's actual/predicted preference over item categories $\vec{c}$	User preference
P, Q	Actual/predicted preferences for users U	Probabilistic distribution of users U 's actual/predicted preference over item categories $\vec{c}$	
$\bar{P}, \bar{Q}$	Typical actual/predicted preferences	Mean actual/predicted preferences P, Q of users U over categories $\vec{c}$	
AT_u	Atypicality	Deviation of a user's preference from typical preference	User characteristics
DV_u	Diversity	Variance of a user preference over item categories $\vec{c}$	
$MC_u(p, q)$	Individual-level miscalibration	Individual-level discrepancy between a user u 's actual preference p and predicted preference q	System error
$MC(P, Q)$	System-level miscalibration	System-level discrepancy between actual preferences P and predicted preferences Q over users U	
$BE_u(p, q)$	Individual-level bias effect	The amount of changes in prediction error for a user u when using the predicted mean preference $\bar{Q}$ instead of actual mean preference $\bar{P}$	System-induced effect
$B(P, Q)$	System-level bias	Systematic distortion in distributions between typical actual/predicted preferences $\bar{P}$	
ST	System-level stereotype	Systematic overgeneralization of predicted preferences Q over users U	
ST_u	Individual-level stereotype	Systematic overgeneralization of a predicted preference q for a user u	
IDV_u	Individual-level inflated diversity	Systematic inflation of a user's predicted preference q for a user u	

Table 1. Mathematical notion, term, definition, and type of concepts in this study.

3.1 Miscalibration

As first introduced in [39], miscalibration quantifies the discrepancy in distributions between an individual's actual preference $p(\vec{c}|u)$ and predicted preferences $q(\vec{c}|u)$ over item categories $\vec{c}$ (e.g., movie genres) for a user u :

- $p(\vec{c}|u)$: The distribution over categories $\vec{c}$ of the set of items interacted by a user (i.e., actual preference).
- $q(\vec{c}|u)$: The distribution over categories $\vec{c}$ of the set of recommended items in a user (i.e., predicted preference).

The individual-level miscalibration, MC_u , for a user u , uses Kullback-Liebert divergence D_{KL} to calculate the discrepancy between two probability distributions to measure how q deviates from p [39], as follows:

$$MC_u(p, q) = D_{KL}(p||q) = \sum_c p(c|u) \log \frac{p(c|u)}{q(c|u)} \quad (1)$$

Miscalibration in the range of $[0, \infty)$ gets smaller when q and p are similar. The zero miscalibration indicates perfect calibration. Similarly, the system-level miscalibration can be denoted as $MC(P, Q)$, where P and Q indicate the sets of actual and predicted preferences for all users.

3.2 Bias-Variance Decomposition for miscalibration

In this section, we present the statistical measure for decomposing miscalibration into bias and variance. We employ the mathematical equation of Bias-Variance decomposition under Kullback-Liebler divergence loss [10, 23], where each user's preferences are represented as a probabilistic distribution over multiple categories of items, and the loss indicates the discrepancy between the distributions of the user's actual and predicted preferences.

System-level bias and variance. First, we show that the miscalibration can be decomposed into bias and variance for all users, which follows the decomposition of Kullback-Leibler divergence as introduced in [9, 23]:

$$MC(P, Q) = \underbrace{D_{KL}(\bar{P}||\bar{Q})}_{\text{Bias}} + \underbrace{\mathbb{E}[D_{KL}(\bar{Q}||Q)]}_{\text{Variance}}. \quad (2)$$

In Eq. 2, two distributions P and Q , which refer to as the data and predictions in the general machine learning setting, can be translated as the set of actual/predicted preferences over users U . In this context, two notations $\bar{P}$ and $\bar{Q}$ indicating the mean of the distributions P and Q can be referred to as the *typical* preference. Miscalibration can be decomposed into two sources of error: (i) **Bias**: the discrepancy between the means of actual preferences and predicted preferences (i.e., the discrepancy between typical preferences). It quantifies the part of error *between* actual and prediction in terms of typical preference. (ii) **Variance**: the total variability of predicted preferences. This quantifies another part of the error *within* predicted preferences, which is how much individual users' predicted preferences deviate from the the typical preference (i.e., mean preference).

Individual-level bias and variance effect. Next, we introduce statistical measures of bias and variance effect at the individual level. Different from the system-level bias and stereotype, this measure quantifies how miscalibration for each user is associated with bias and variance at the individual level. Inspired by the decomposition of general loss into bias-variance effect [21, 23], we propose individual-level bias and variance effects under KL-divergence loss as follows:

$$MC_u(p, q) = \underbrace{(D_{KL}(p||\bar{Q}) - D_{KL}(p||\bar{P}))}_{\text{Bias effect}} + \underbrace{(D_{KL}(p||q) - D_{KL}(p||\bar{Q}))}_{\text{Variance effect}}, \quad (3)$$

where p, q indicate the actual and predicted preferences for an individual user u . Following the definitions in [23], individual-level miscalibration can be decomposed into two terms defined as follows : (i) **Individual bias effect**: the amount of changes in prediction error for a user u when using the predicted mean preference (i.e., $\bar{Q}$) instead of the actual mean preference (i.e., $\bar{P}$). (ii) **Individual variance effect**: the discrepancy between prediction errors with and without variability in a given prediction. More specifically, a prediction q can differ across users, whereas there is no variability if all users receive the same prediction according to the mean $\bar{Q}$.

Distinction between system-level effect (Eq. 2) and individual-level effect (Eq. 3). In contrast to the miscalculation decomposed at the collective level (Eq. 2), Eq. 3 specifically looks into the bias and variance decomposition of the individual-level miscalculation by introducing a user's actual and predicted preferences, namely p and q . The specification follows a difference-in-differences paradigm: The individual bias effect measures the difference in the differences between a user's actual preference and the mean actual preferences and between the user's actual preference and the mean predicted preferences. The individual variance effect, on the other hand, measures the difference in the differences between a user's actual and predicted preferences, and between the user's actual preference and the mean predicted preferences. Both can take a positive or negative value, which contributes to the increasing or decreasing miscalibration for a given user.

3.3 System-induced effects

Compared to Section 3.2, the measures on system-induced effects presented in this section highlight two types of quantities of errors relating to bias and variance respectively.

System-level bias and stereotype. At aggregated level, two system-level effects quantify the overall errors introduced by the model across all users.

- **System-level bias:** Systematic distortion between mean actual and predicted preferences. This is equivalent to overall bias in Eq. 2 that can be denoted as $B(P, Q) = D_{KL}(\bar{P}||\bar{Q})$.
- **System-level stereotype:** Systematic overgeneralization of predicted preferences based on certain characteristics of individuals, defined as follows:

$$ST(P, Q) = 1 - \frac{\mathbb{E}_{q \in Q}[D_{JS}(q||\bar{Q})]}{\mathbb{E}_{p \in P}[D_{JS}(p||\bar{P})]}, \quad (4)$$

where the mean deviations of actual and predicted preferences P and Q for users U from typical preference $\bar{P}, \bar{Q}$ are computed as $\mathbb{E}_{p \in P}[D_{JS}(p||\bar{P})]$ and $\mathbb{E}_{q \in Q}[D_{JS}(q||\bar{Q})]$ respectively. We use Jensen-Shannon divergence to compute the deviation for individual user's preference p from typical preference $\bar{P}$, where $D_{JS}(p||\bar{P}) = \frac{D_{KL}(p||\bar{P}) + D_{KL}(\bar{P}||p)}{2}$ is a symmetric measure of quantifying the distance between two probabilistic distributions. The system-level stereotype measures the relative difference between two quantities, i.e., how the variance in actual preferences decreased in predicted preferences, which has a range of $(-\infty, 1)$. When $ST(P, Q) > 0$, it indicates user preferences in recommendations stereotyped towards typical preference at the system-level. The zero quantity in $ST(P, Q)$ indicates no overall stereotypes.

Individual-level stereotype. The stereotype at the individual level is defined as follows:

$$ST_u(p, q) = D_{JS}(p||\bar{P}) - D_{JS}(q||\bar{Q}). \quad (5)$$

Individual-level inflated diversity. We introduce the concept of "inflated diversity" to describe the situation in which a user receives an over-diverse prediction q , i.e., one that is more diverse than the user's actual preference p . It quantifies how a user with particular interests is forced to resemble dominant groups with diverse preferences.

$$IDV_u(p, q) = DV_u(p) - DV_u(q), \quad (6)$$

where DV_u measures the diversity of a user u (see Eq. 9).

Bias disparity. Introduced by [41], bias disparity measures the relative difference in the preference ratio of users U on an item category c between actual and predicted preference.

$$BD(c, U) = 1 - \frac{q(c, U)}{p(c, U)}, \quad (7)$$

where $p(c, U)$ and $q(c, U)$ denote the distribution of actual and predicted preferences over an item category c of the set of items for users U . When $BD(c, U) > 0$, it indicates the amplification of preference ratio on item category c in predicted preference compared with actual preference. We leverage this measure to interpret system-induced effects at category level.

3.4 User characteristics

We further consider two types of user characteristics, atypicality and diversity, which are related to the variability of user preferences over item categories.

Atypicality. This quantifies the deviation of a user's preference from typical preference, which can be measured by the distance between two distributions, typical preference and a user's preference:

$$AT_u = D_{JS}(p||\bar{P}). \quad (8)$$

Similarly, the predictive atypicality in the range of $[0, 1]$ can be defined by replacing the observed quantities p and $\bar{P}$ with the two predicted quantities q and $\bar{Q}$, respectively.

Diversity. This measures the variance of individual preferences:

$$DV_u = Entropy(u) = - \sum_{c \in \vec{c}} p(\vec{c}|u) \log p(\vec{c}|u), \quad (9)$$

where *Entropy* denotes the normalized entropy of preference distribution for an individual user u , which has a range of $[0, 1]$. A higher diversity indicates a user preference varying over multiple categories to a similar degree. A lower diversity indicates that a user prefers certain categories particularly more than others.

4 ANALYSIS RESULTS

Based on the proposed measurement framework, we investigate the miscalibration, system-induced effects, and disparate impact of recommender systems. We examine algorithmic, group, and individual differences in the context of movie recommendation in order to answer the following three analytic questions:

- **RQ1.** To what extent do two distinct sources of error, bias, and variance, contribute to miscalibration?
- **RQ2.** What is the impact of stereotypes and biases induced by a recommender system on different individuals and groups?
- **RQ3.** What is the relationship between user characteristics, systematic effects, and miscalibration?

4.1 Experimental setting

Data. In the analysis, we use MovieLens 1M dataset¹ that contains demographic attributes such as gender and age. The original dataset includes 1,000,209 user ratings ranging from 1 to 5 over movie items across 18 movie genres. We take three preprocessed steps frequently adopted in existing work [3, 27, 43] to follow the best practice and ensure our work comparable to them: (1) We binarized ratings by dropping all user-item records with ratings less than 4 to convert it to implicit feedback. Then, (2) we remove all users with less than 20 interactions to ensure the quality of preference information of each individual. Lastly, (3) for items associated with more than multiple genres, we normalized scores in each item to be proportional to the number of per-category multiple categories. After the preprocessing steps, the data includes 562,800 interactions from 5,180 users and 3,526 items.

Following the experiment settings in previous literature [3, 39], the implicit feedback data was split into 80% and 20% for training and testing using user-fixed setting in Librec-auto package where the interactions within each user were divided into training and test set with respect to the ratio based on chronological order. Experiments for all five algorithms (details below) were conducted with the same train-test split data. We employ the same setting to facilitate the comparison of all results derived from predicted recommendations.

¹<https://grouplens.org/datasets/movielens/1m/>

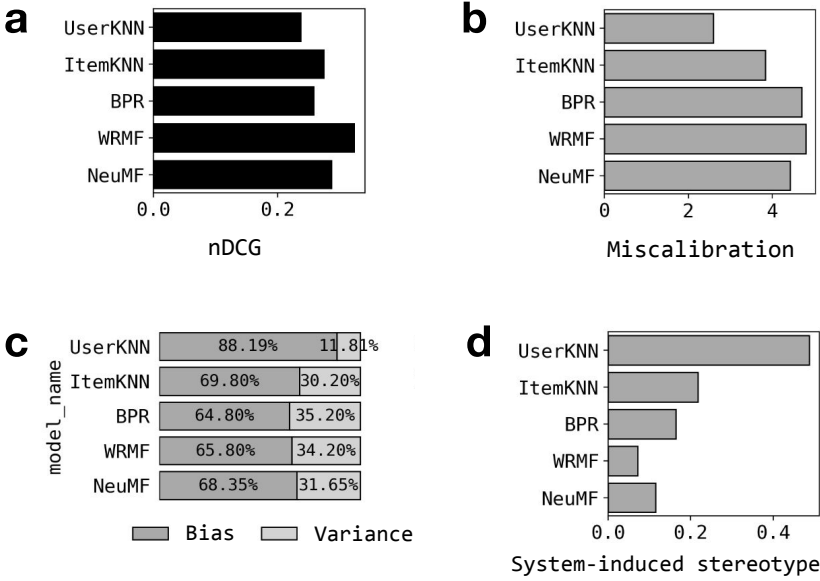


Fig. 2. The algorithmic difference of system-induced effects and recommendation quality over five recommendation algorithms. Two performance results in (a) nDCG@20 and (b) miscalibration@20 show that complex models tend to perform better for item-wise predictions (nDCG@20) but worse at category-wise calibration (miscalibration@20). (c) In Bias-Variance ratio, complex models tend to have a higher variance, indicating the capability of better approximating the variance in actual preference. (d) This in turn leads to lower system-induced stereotype.

Algorithms. In the experiment, we compare five recommendation algorithms to examine the generalizability of our observation on system-induced effects. Our choice of algorithms, based on a survey on related work in line of research, is to ensure a variety of recommendation algorithms to be examined, encompassing four types of methods ranging from the baselines to state-of-art algorithms, including memory-based (ItemKNN [29], UserKNN [38]), learning-to-rank (Bayesian Personalized Ranking, BPR [37]), latent factor (Weighted Regularized Matrix Factorization, WRMF [22]) and deep learning model (Neural Matrix Factorization, NeuMF [20]). These five recommender systems were experimented with hyperparameters with the highest performance based on the hyperparameter search supported in the Librec-auto package (UserKNN, ItemKNN, BPR, and WRMF using Librec-auto² [31, 32] and NeuMF using QRec³ for deep-learning model [47], which are python-based framework for recommender systems). The algorithms were evaluated with Normalized Discounted Cumulative Gain (nDCG) [24] and miscalibration (Eq. 1) for top 20 recommended items.

4.2 RQ1. Algorithmic differences in miscalibration and system-induced effects

We first analyze the algorithmic differences in miscalibration, system-induced bias, and stereotype, with the goal of understanding how complex and simple models behave differently. Among the five algorithms tested, WRMF and NeuMF are considered more complex models than the others. Fig. 2 summarizes the results of algorithmic differences, including: the recommendation performances at

²<https://github.com/that-recsys-lab/librec-auto>

³<https://github.com/Coder-Yu/QRec>

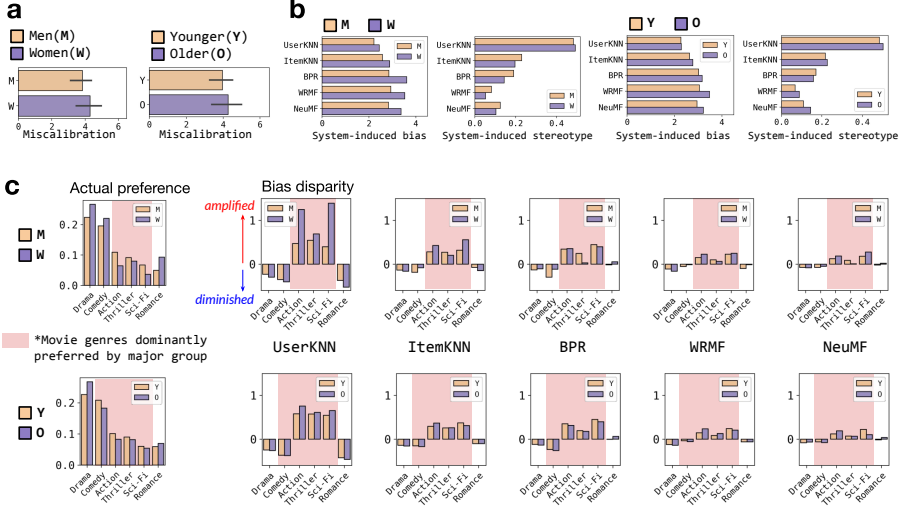


Fig. 3. Disparate impact of miscalibration and system-induced effects over groups. (a) Miscalibration: For both demographic attributes, gender and age, five recommendation algorithms on average exhibit higher miscalibration for minority groups, women and older users. (b) System-induced effects: Both bias and stereotype tend to impact minority groups to a greater extent. (c) Bias disparity: At the category-level, user preferences in majority-dominant genres (i.e., categories dominantly preferred by majority group) highlighted as red areas tend to be amplified.

the item level (in terms of nDCG, Fig. 2a) and the item-categorical level (miscalibration, Fig. 2b), Bias-Variance ratio as sources of miscalibration (Fig. 2c), and the system-induced stereotype (Fig. 2d).

The results indicate that complex models, such as WRMF or NeuMF, tend to perform better for item-level predictions but less well for category-level calibration of user preferences. The two models, UserKNN and NeuMF, are two examples of relatively simple (memory-based) and complex (deep learning method) approaches, respectively. We have three key observations. First, the simplest model, UserKNN, had the lowest performance in nDCG but the least amount of miscalibration (Fig. 2a-b). Second, simpler models have a smaller proportion of variance (Fig. 2c), suggesting that the ratio of variance is positively associated with model complexity. In particular, UserKNN had the lowest variance ratio of the five algorithms examined. Lastly, system-induced stereotypes tend to be more prevalent in simpler models (Fig. 2d). UserKNN exhibited the highest degree of stereotyping, whereas others exhibited a lower degree of stereotyping.

These findings are consistent with a well-known bias-variance trade-off [26] and its association with model complexity, which also applies to recommender systems: Simple recommendation algorithms are more susceptible to bias, whereas complex models have greater variance. Our analysis also reveals that a higher level of predictive variance is positively correlated with a lower level of stereotype. When a model's predictive preferences more closely resemble the variability of users' actual preferences, the impact of stereotyped predictions on users is reduced.

4.3 RQ2. Disparate impact of miscalibration, bias, and stereotype over groups and individuals

We investigate how miscalibration and system-induced effects affect different groups and users disproportionately. We use the statistical measures presented in Section 3 to measure the collective- and individual-level effects in terms of miscalibration, system-induced bias and stereotype.

4.3.1 Group-level disparity over demographic groups. First, we explore the group-level disparity between demographic groups, focusing on two demographic attributes, gender and age (see details for demographic attributes in Section 4.1).

The disparate impacts of miscalibration and two system-induced effects for both demographic attributes are summarized in Fig. 3. For miscalibration, two minority groups, women and older users, have a higher degree of miscalibration than majority groups in average over five algorithms (Fig. 3a), with a significant difference validated in t -test ($p \leq 0.001$). Regarding system-induced effects (Fig. 3b), while system-induced bias showed higher impact on minority groups in both gender and age regardless of algorithms, system-induced stereotype had a mixed result. For the age attribute, older users had a higher level of stereotypes across all algorithms. For gender, on the other hand, women or men users interchangeably had a greater stereotype in different algorithms. For the group disparity at the category level (Fig. 3c), we observe a consistent pattern of bias disparity (details in Section 3.3), which tends to amplify the preference of majority-dominant categories (i.e., categories dominantly preferred by majority group) such as action, thriller, and romance as men-dominant genres and comedy, action, thriller, and sci-fi as genres dominated by younger users (highlighted as red in Fig. 3c). However, the degree of category-level bias disparity varies across algorithms. Simpler algorithms tend to show higher bias disparity in majority-dominant genres and in minority groups to a greater extent.

The bias disparity tends to increase in majority-dominant categories, such as action, thriller, and romance genres favored predominantly by men, and comedy, action, thriller, and sci-fi genres favored predominantly by younger users. In Fig. 3c, these categories with a majority dominance are highlighted in red. On the other hand, the degree of the category-level bias disparity varies across algorithms. Simpler models tend to exhibit a greater bias disparity in majority-dominant genres and among minority groups.

4.3.2 Individual-level disparity between typical vs. atypical users. Next, we investigate how these system-induced effects impact typical vs. atypical users disproportionately. This allows us to see how the degree of users' preferences deviated from the average preference can be associated with disparate impacts of system-induced effects and miscalibration.

Fig. 4 shows the individual-level system-induced effects, including (from left to right) miscalibration, bias and variance effects, stereotyping, and inflated diversity, where users are sorted according to their typicality (top: the most typical users, bottom: the most atypical users). The two curves represent the trends of system-induced effects over user typicality for different age (Fig. 4a) and gender (Fig. 4b) groups.

We see a clear difference between typical and atypical users in terms of the impact of stereotype and inflated diversity. First, while there is no discernible difference between typical and atypical users in terms of miscalibration, the bias and variance effect impact these user groups differently. In particular, atypical users have a greater bias effect and a lower variance effect than typical users. Second, stereotyping and inflated diversity, two system-induced effects at the individual level, tend to be positively associated with atypicality. Users with more specific preferences at the bottom of the plots, as shown in Fig. 4, tend to be more stereotyped and have a higher degree of inflated diversity. In contrast, typical users tend to experience the opposite effect, with their preferences

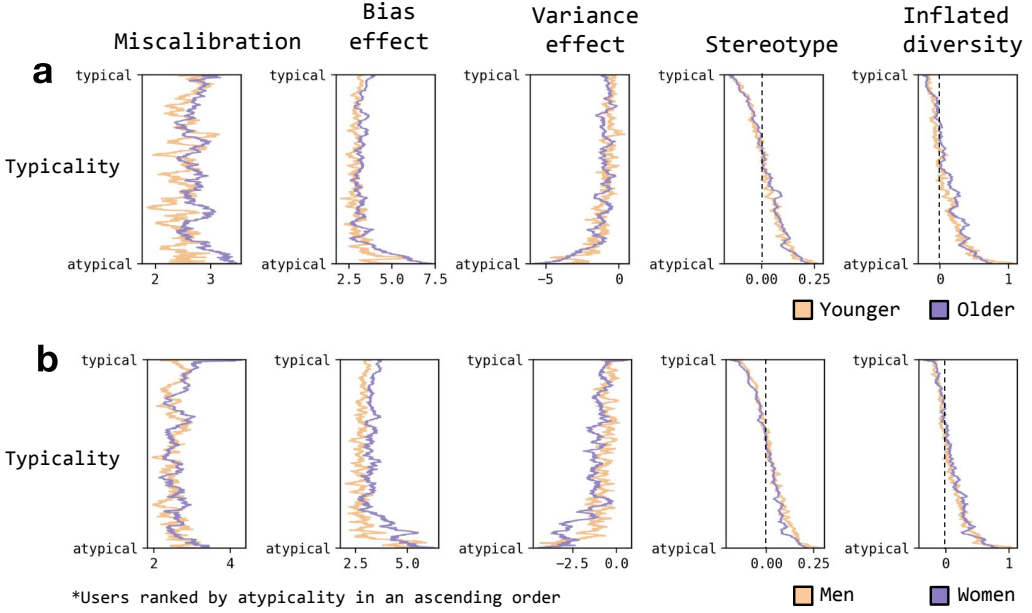


Fig. 4. The individual-level impact of miscalibration, bias/variance effect, stereotype, and inflated diversity in WRMF algorithm for (a) gender and (b) age groups. All individuals are ranked by atypicality in order of the most typical (top) to the most atypical (bottom) users and distinguished by gender (younger and men users in orange; older and women users in purple).

becoming inversely stereotyped and less diverse. This indicates that the recommendations received by atypical users tend to be overgeneralized as a result of the combined effects of stereotyping and inflated diversity.

4.4 RQ3. Relationship between user characteristics, system-induced effects, and miscalibration

In the previous sections, we discovered that recommendation algorithms consistently have a disparate impact on groups and individuals, even though they exhibit miscalibration and system-induced effects to varying degrees. As a result, minority groups and atypical individual users were discriminated against. Given all of the findings from RQ1 and RQ2, we intend to systematically investigate the relationship between user characteristics, system-induced effects, and miscalibration.

We hypothesize that a recommender system's miscalibration (between users and their preferred genres) is a combined result of the various forms of bias produced by its model, trained with data comprising different kinds of users. Fig. 5 illustrates our hypothesis. We consider two user factors that can be directly measured from the input data (**Data** phase), atypicality and diversity, four types of system-induced effects (individual-level bias, variance, stereotype, and inflated diversity) that are produced by a model (**Model** phase), and the prediction performance evaluated in terms of miscalibration (**Prediction** phase). We use Structural Equation Modeling (SEM) [42] to statistically test the associations between different factors, as it allows us to test the direct and indirect relationships among the factors.

The result of structural equation modeling is summarized in Fig. 5. All paths are annotated with its standardized coefficient of the mean and 95% confidence interval from SEMs across five

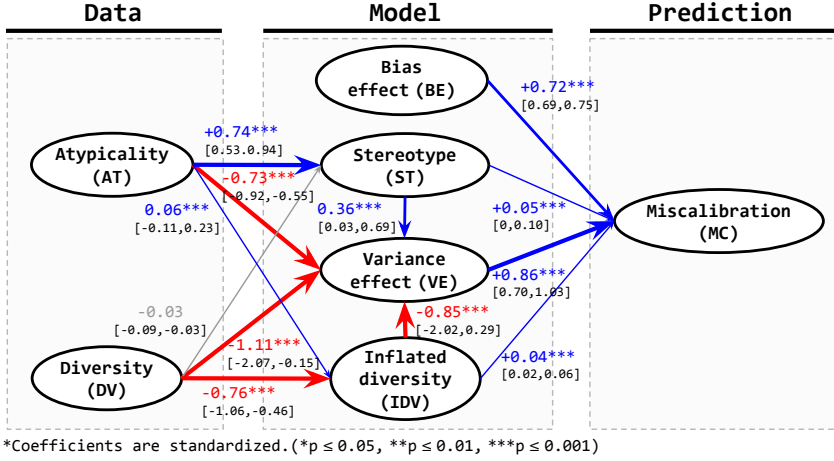


Fig. 5. The relationship between systematic effects and miscalibration across five recommendation algorithms.

recommendation algorithms. The SEM results revealed several interesting findings. First, two user characteristics, atypicality and diversity, are negatively associated with each other (Fig. 6a). This means that typical users tend to prefer diverse movie genres while atypical users tend to have particular interest in certain genres. Second, atypicality and diversity have strong associations with stereotype and inflated diversity respectively: atypicality is positively associated with stereotype indicating atypical users to be more stereotyped; diversity is negatively associated with inflated diversity, meaning that user preferences with a particular interest tend to be forced to be diverse.

To further look into the relationship between two user characteristics, we divide users in four groups by atypicality (low/high) and diversity (low/high), categorized by thresholding the measures using median value. These four groups of users are visually differentiated in Fig. 6b, where all users in MovieLens 1M dataset were projected into two-dimensional space using UMAP (Uniform Manifold Approximation and Projection) [34] based on their genre-wise preferences. In the plot, the users with typical and diverse preference (colored as orange) is placed at the center of the user space, while users with atypical and less diverse preference tend to be placed at the periphery. We also find that these four groups exhibit significantly different degree of stereotype and inflated diversity (one-way ANOVA; $p \leq 0.001$). As shown in Fig. 6b, users with atypical and less diverse (AT/NDV) are the most stereotyped and have inflated diversity, while users with typical/diverse are rather inverse-stereotyped and have reduced diversity, indicating the negative association between intersectional user characteristics and two system-induced effects.

Third, two system-induced effects, stereotype and inflated diversity, were found to have the opposite association with variance effect: higher stereotype tends to increase variance effect (ST → VE; 0.36 ± 0.17) while inflated diversity tends to decrease variance effect (IDV → VE; -0.85 ± 0.60). From model to prediction phase, variance effect is the most highly associated with miscalibration (VE → MC; 0.86 ± 0.10). Lastly, indirect effects from stereotype and inflated diversity to miscalibration also exist. Higher stereotype tends to increase miscalibration (ST → VE → MC; 0.3 ± 0.1) while higher inflated diversity tend to decrease miscalibration (IDV → VE → MC; -0.71 ± 0.38), indicating they have an indirect effect on miscalibration.

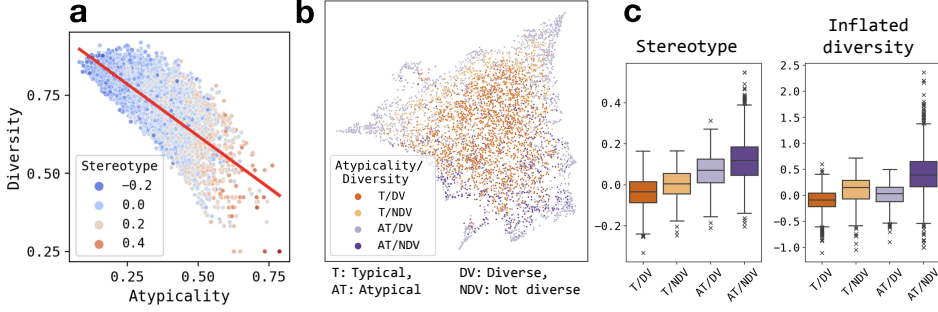


Fig. 6. The relationship between systematic effects and miscalibration across five recommendation algorithms. (a) Atypicality and diversity are negatively associated with each other. This indicates that atypical users tend to have particular and less diverse preferences. (b) Four groups of users based on their atypicality and diversity are clearly distinct over user space, confirming the relationship between users' deviation from and two user characteristics. (c) Four groups of users are significantly different in two system-induced effects. More atypical and diversity users are found to have higher impact of stereotype and inflated diversity.

5 MITIGATING UNDERREPRESENTATION PROBLEM

Throughout R1-R3, we observed that recommendation algorithms tend to exhibit system-induced effects and miscalibration to a different extent but consistently have disparate impacts on minority groups such as women and older users. Bias disparity was also observed to amplify preferences in majority-dominant genres in both gender and age groups. At the individual level, atypical users tended to be stereotyped and have their diversity inflated to a greater extent. All of these findings indicate the systemic effects of overgeneralizing user preferences toward typical/dominant preferences, which impact users who are underrepresented in the data.

Based on these findings, we examine the possibility of mitigating the disparate impact on the underrepresented users in the data. We hypothesize that modifying the proportion of underrepresented users in the training data will change algorithmic behavior. Thus, our strategy is to use bootstrapped resampling to adjust the data distribution (users' interactions with items) in order to increase the data representation of stereotyped users. We use an equal-sized binning that divides users into n bins as follows. First, we determine the number of bins n where each bin has an equal number of users ($1/n$ of the total number of users). These bins are further split into two groups: the stereotyped ($ST_i > 0$ from Eq. 5) who will be oversampled and the inverse-stereotyped ($ST_i < 0$ from Eq. 5) who will not be involved in oversampling into different bins. We determine n such that the inverse-stereotyped users are contained in a single bin, and the remaining bins are the stereotyped users.

After binning, we determine the oversampling rates $R_k = \{r_k, 1 + b * k\}_{k \in N}$ for stereotyped users, where these rates for n bins are set to linearly increase based on a base sampling rate b between 0.01 to 0.4. Based on oversampling rates, individual users' interaction for each bin is sampled with replacement based on the rate. For oversampling rates greater than 1, we sample only for the portion of a rate ($r_k - 1$) and then combine the sampled data with the the original data as the portion of 1. After resampling, we train five recommendation models with the updated dataset. We report the results from the original and oversampled training data in Table 2 that shows the best performance among different base oversampling rates in terms of nDCG@20, Stereotype, and Miscalibration@20.

The results show that oversampling stereotyped users overall helps improve recommendation quality, miscalibration, and stereotype, with its effectiveness over different recommendation models.

	Original				Oversampling			
	% ST-ed users	nDCG @20	ST	MC @20	nDCG @20	ST	MC @20	Base sampling rate
UserKNN	93.5%	0.2390	0.4708	2.7946	0.2506	0.4667	2.8116	0.4
ItemKNN	74.5%	0.2751	0.2151	3.8714	0.2765	0.2373	3.8609	0.15
BPR	66.4%	0.2598	0.1643	4.6925	0.2602	0.1596	4.8234	0.15
WRMF	55.9%	0.3251	0.0714	4.7914	0.3258	0.1023	4.7618	0.1
NeuMF	44.2%	0.2885	0.1156	4.4244	0.2606	0.0617	5.3689	0.05

Table 2. The metrics of before (original) and after oversampling stereotyped users.

Specifically, simpler models such as UserKNN or ItemKNN with the highest stereotypes from a larger pool of users have the largest gain of nDCG. On the other hand, NeuMF, the most complex model with the lowest stereotype, has an adverse effect with its decreasing nDCG. We also find that a greater degree of oversampling (i.e., larger base sampling rates) works better for models with greater stereotype.

While the oversampling technique generally improves stereotyping and miscalibration, its effect varies widely across algorithms and is inconsistent between two measures. When stereotype increases or decreases, miscalibration has the opposite effect for all algorithms. This poses an additional challenge to the effectiveness of mitigating user underrepresentation: system-induced effects and miscalibration have complex interactions in model results.

6 LIMITATION AND FUTURE WORK

In this section, we discuss the limitations of our work, implications and future work in tackling the problem of system-induced effects in recommender systems.

Deeper understanding of user characteristics, system-induced effects, and miscalibration.

Our study presents a grounded work to intricately explore and establish connections between various factors in different stages of recommender systems, among user characteristics, system-induced effects and errors. However, we find that the scope of the current study is limited in several ways. First, while our analysis primarily investigated two systematic effects – stereotype and inflated diversity – associated with variance effects, we have yet to examine associations related to other factors, such as other system-induced filter bubble or echo chamber effects [18, 36], and their relationship with errors. Those effects are somewhat related to our observations, especially the effect of bias amplification exaggerating preferences over the major categories in our observations, but might be further explained in relation to errors and its components, bias and variance, as examined using our framework. Second, conducting an intersectional analysis across subgroups (e.g., black women) and their category-level preferences may reveal unknown associations, shedding light on the degree to which those subgroups are likely to face system-induced effects, and its relationship with their diversity and atypicality. Lastly, the alignment of system-induced stereotypes with social stereotypes also remains underexplored. Only a few studies have investigated the discrepancy between actual gender difference and social stereotype in some domains such as movie genres [45]. Such analysis could unveil how data and algorithmic recommendation encode and perpetuate stereotypes minded among people into automatic systems.

Mitigation strategies. To explore the possibility of mitigating system-induced effects, we experimented with an oversampling method – an approach to overrepresent users who were stereotyped in the dataset. While we used stereotype scores as a threshold for oversampling, the choice of score

may depend on how underrepresented users are defined, considering factors like inflated diversity or user characteristics (diversity or typicality).

Despite its overall effectiveness, we observed the mixed effect of our mitigation strategy (i.e., the trade-off between miscalibration and stereotype), indicating the complex relationship between system-induced effects. This could be partially attributed to a challenging nature of mitigating variance effect, indicating that highly varied individual preferences may not be able to be corrected via automated approach. Given this challenge, a user-driven mitigation, allowing users to review miscalibrated and stereotyped items with explanations and interface for intervention, can be a potential direction to address the gap. This approach can further ensure better transparency with a greater degree of user autonomy.

In addition, although our oversampling method effectively mitigates various types of errors and enhance data representation, the approach needs to be used with caution. For instance, it may lead to overfitting towards minority users whose instances are sampled multiple times, resulting in greater model bias, and poor performances on unseen data.

Dichotomous analysis. Our analysis was conducted in binary group settings due to limited data availability (the original dataset including two genders and limited point of ages). A further analysis for multiple groups with available data may lead to different results, which may help generalize our findings to other demographic attributes such as other gender groups or other demographic groups such as ethnicity.

Application scope. Also, future work in different recommendation contexts and more advanced algorithms can also help make our findings more generalizable to a wide range of recommender systems. This may include a diverse type of recommendation ranging from products, online news, and music to insurance and job application.

7 CONCLUSION

In this study, we proposed a unified framework for examining stereotype, bias, and miscalibration. Our framework demonstrated that the problem of miscalibration discussed in previous studies [2, 3, 14, 39] can be better understood by Bias-Variance decomposition, which can then be examined further through two system-induced effects, bias and stereotype, respectively. In the analysis of movie recommendation using the framework, we showed that recommendation algorithms exhibit systematic behaviors differently in bias and variance, but consistently introduce a higher degree of system-induced bias and stereotype in minority groups and atypical users, and amplify the preference ratio of majority-dominant genres. All of these observations illustrate the systematic tendency for underrepresented users' preferences to be overgeneralized in favor of typical preferences. Our mitigation strategy, the oversampling of stereotypical users, was found to not only enhance recommendation quality but also mitigate the system-induced stereotypes.

ACKNOWLEDGMENTS

REFERENCES

- [1] Himan Abdollahpour, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 42–46.
- [2] Himan Abdollahpour, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Impact of Popularity Bias on Fairness and Calibration in Recommendation. <http://arxiv.org/abs/1910.05755> arXiv:1910.05755 [cs].
- [3] Himan Abdollahpour, 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*. ACM, Virtual Event Brazil, 726–731. <https://doi.org/10.1145/3383313.3418487>

- [4] Nourah A ALRossais and Daniel Kudenko. 2018. Evaluating Stereotype and Non-Stereotype Recommender Systems.. In *KaRS@ RecSys*. 23–28.
- [5] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
- [6] Shilong Bao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. The Minority Matters: A Diversity-Promoting Collaborative Metric Learning Algorithm. <http://arxiv.org/abs/2209.15292> arXiv:2209.15292 [cs].
- [7] Joeran Beel, Siddharth Dinesh, Philipp Mayr, Zeljko Carevic, and Jain Raghvendra. 2017. *Stereotype and most-popular recommendations in the digital library sowiport*. Humboldt-Universität zu Berlin.
- [8] Lawrence Blum. 2004. Stereotypes And Stereotyping: A Moral Analysis. *Philosophical Papers* 33, 3 (Nov. 2004), 251–289. <https://doi.org/10.1080/05568640409485143>
- [9] James Brofos, Rui Shu, and Roy R Lederman. [n. d.]. A Bias-Variance Decomposition for Bayesian Deep Learning. ([n. d.]).
- [10] James Brofos, Rui Shu, and Roy R Lederman. 2019. A bias-variance decomposition for Bayesian deep learning. In *NeurIPS 2019 Workshop on Bayesian Deep Learning*.
- [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*. ACM, Las Vegas Nevada, 1–8. <https://doi.org/10.1145/1722149.1722154>
- [12] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, Vancouver British Columbia Canada, 224–232. <https://doi.org/10.1145/3240323.3240370>
- [13] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to Recommend Accurate and Diverse Items. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Perth Australia, 183–192. <https://doi.org/10.1145/3038912.3052585>
- [14] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 172–186. <https://proceedings.mlr.press/v81/ekstrand18b.html> ISSN: 2640-3498.
- [15] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, Vancouver British Columbia Canada, 242–250. <https://doi.org/10.1145/3240323.3240373>
- [16] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (Nov. 2020), 102377. <https://doi.org/10.1016/j.ipm.2020.102377>
- [17] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding Echo Chambers in E-commerce Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event China, 2261–2270. <https://doi.org/10.1145/3397271.3401431>
- [18] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 2261–2270.
- [19] Wenshuo Guo, Karl Krauth, Michael Jordan, and Nikhil Garg. 2021. The Stereotyping Problem in Collaboratively Filtered Recommender Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, – NY USA, 1–10. <https://doi.org/10.1145/3465416.3483298>
- [20] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [21] Tom Heskes. 1998. Bias/Variance Decompositions for Likelihood-Based Estimators. *Neural Computation* 10, 6 (Aug. 1998), 1425–1433. <https://doi.org/10.1162/089976698300017232>
- [22] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.
- [23] Gareth M James. 2003. Variance and bias for general loss functions. *Machine learning* 51 (2003), 115–135.
- [24] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
- [25] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 3819–3828. <https://doi.org/10.1145/2702123.2702520>

- [26] Ron Kohavi, David H Wolpert, et al. 1996. Bias plus variance decomposition for zero-one loss functions. In *ICML*, Vol. 96. Citeseer, 275–283.
- [27] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2019. Crank up the volume: preference bias amplification in collaborative recommendation. <http://arxiv.org/abs/1909.06362> arXiv:1909.06362 [cs, stat].
- [28] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2020. Calibration in Collaborative Filtering Recommender Systems: a User-Centered Analysis. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. ACM, Virtual Event USA, 197–206. <https://doi.org/10.1145/3372923.3404793>
- [29] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.
- [30] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. 2020. Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems. <http://arxiv.org/abs/2002.07786> arXiv:2002.07786 [cs].
- [31] Masoud Mansoury and Robin Burke. 2019. Algorithm Selection with Librec-auto.. In *AMIR@ ECIR*. 11–17.
- [32] Masoud Mansoury, Robin Burke, Aldo Ordóñez-Gauger, and Xavier Sepúlveda. 2018. Automating recommender systems experimentation with librec-auto. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 500–501.
- [33] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. 2019. Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison. <http://arxiv.org/abs/1908.00831> arXiv:1908.00831 [cs].
- [34] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [35] An-Te Nguyen, Nathalie Denos, and Catherine Berrut. 2007. Improving new user recommendations with rule-based induction on cold user data. In *Proceedings of the 2007 ACM conference on Recommender systems*. 121–128.
- [36] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*. ACM, Seoul Korea, 677–686. <https://doi.org/10.1145/2566486.2568012>
- [37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
- [38] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 175–186.
- [39] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, Vancouver British Columbia Canada, 154–162. <https://doi.org/10.1145/3240323.3240372>
- [40] Cass R Sunstein. 2009. *Going to extremes: How like minds unite and divide*. Oxford University Press.
- [41] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. 2018. Bias Disparity in Recommendation Systems. *arXiv:1811.01461 [cs]* (Nov. 2018). <http://arxiv.org/abs/1811.01461> arXiv: 1811.01461.
- [42] Jodie B Ullman and Peter M Bentler. 2012. Structural equation modeling. *Handbook of Psychology, Second Edition 2* (2012).
- [43] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded Recommendation for Alleviating Bias Amplification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Aug. 2021), 1717–1725. <https://doi.org/10.1145/3447548.3467249> arXiv: 2105.10648.
- [44] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable Recommendation Against Filter Bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Madrid Spain, 1251–1261. <https://doi.org/10.1145/3477495.3532075>
- [45] Peter Wüth, Benjamin P. Lange, and Sascha Schwarz. 2017. Tears or Fears? Comparing Gender Stereotypes about Movie Preferences to Actual Preferences. *Frontiers in Psychology* 8 (March 2017). <https://doi.org/10.3389/fpsyg.2017.00428>
- [46] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the Long Tail Recommendation. <http://arxiv.org/abs/1205.6700> arXiv:1205.6700 [cs].
- [47] Junliang Yu, Hongzhi Yin, Min Gao, Xin Xia, Xiangliang Zhang, and Nguyen Quoc Viet Hung. 2021. Socially-aware self-supervised tri-training for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2084–2092.
- [48] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*. ACM Press, Chiba, Japan, 22. <https://doi.org/10.1145/1060745.1060754>

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009