From Monolith to Mosaic: Uncovering Behavioral Differences for Choice Models in Recommender Systems Simulations

Robin Ungruh
R.Ungruh@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Alejandro Bellogín
alejandro.bellogin@uam.es
Universidad Autónoma de Madrid
Madrid, Spain

Maria Soledad Pera
M.S.Pera@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Abstract

Simulation is widely used in recommender systems research to study algorithm behavior and its impact on users. A common strategy involves adopting a *universal* choice model to represent users, assuming all follow the same consumption patterns. This one-size-fits-all approach overlooks the diversity in user preferences and decision-making patterns. In this work, we scrutinize whether this universal view fails to account for unique user behavior, thus harming realism and reliability of simulation outcomes. We conduct multiple simulations with various recommendation algorithms and choice models in the movie domain, comparing outcomes to users' organic consumption patterns. Further, we evaluate whether a holistic model that captures users' differences in behavior would better reflect a wide user base. Our findings highlight the limitations of using a naive, universal choice model and emphasize the need for more nuanced, user-specific approaches to make contributions from simulation studies more reflective of real-world effects.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Recommender Systems, Simulations, Choice Models

ACM Reference Format:
Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. From Monolith to Mosaic: Uncovering Behavioral Differences for Choice Models in Recommender Systems Simulations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3726302.3730199

1 Introduction

The use of simulations to, among other purposes, study the behavior of Recommendation Algorithms (RAs) over multiple iterations has gained popularity among Recommender System (RS) research [14]. Simulations enable in-depth analyses of recommendations, revealing changes in the characteristics of recommended items that may only become apparent after repeated interactions [9, 13, 48]. With this approach, RS research aims for more realistic perspectives on real-world RA effects (e.g., amplification of popularity bias [37] and gender bias [21], homogenization [10], and decreased diversity [30]) and long-term consequences for users of RS [16, 18, 29, 49].

RS simulation frameworks rely on assumptions about user behavior, expressed through choice models that govern how users select and consume recommended items, for example, based on a ranked probability, where items ranked higher are more likely to be consumed [e.g., 19, 24, 34, 48]. The selection of a choice model considerably affects what is simulated to be consumed by the user [19, 24, 26]; it directly shapes the conclusions drawn from the simulations as different choice models may lead to simulated choices that reflect users' actual preferences to varying degrees [10, 37, 49].

Realistic choice models are crucial for simulations to capture real-life impacts [9]. Yet, usual choice models tend to simplify user-recommendation interactions [13, 25] and studies rarely validate them against actual consumption patterns, e.g., those captured in RS datasets. This risks that findings based on 'wrong' choice models may be less reflective of real RS scenarios [10, 26]. This concern is exacerbated by the typical adoption of a *universal choice model*, one that applies to all users across all iterations [e.g., 18, 21, 30, 34]. This one-size-fits-all assumption to modeling users neglects heterogeneous behavior [9], including varying preferences, decision-making patterns, and engagement [5, 32]. While simulations under this assumption reveal overall trends, we argue that they may lead to less realistic outcomes and obscure the disproportionate impact of effects like algorithmic biases on some users [37].

We address this research question (RQ): *To what extent does applying universal choice models in recommender system simulation capture the behavioral nuances of the entire user base?* We seek to uncover if a universal model fails to capture the complexities of a wide RS user base, directly addressing the concern that common simplifications result in misaligned depictions of real-world dynamics. For this, we conduct a simulation study applying a simulation pipeline to various combinations of RAs and universal choice models in the movie domain. Instead of merely assessing whether simulated items are relevant to users or have certain desirable properties such as diversity or utility [10, 30], we evaluate misalignments between users' simulated choices and natural consumption patterns over an extended period for each individual. This allows us to gauge whether certain choice models reliably capture some users while failing to reflect others. Further, we analyze whether a *holistic* choice model, one that considers a unique choice model for each user, would better capture the needs of a wide user base.

Findings spotlight the limitations of using the same choice model, as simulated behavior markedly deviates from organic patterns. Even the seemingly *best* model fails to properly capture the needs of a non-negligible number of users. Instead, a holistic choice model improves alignment between simulated and organic consumption

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730199

patterns, highlighting the importance of capturing individual differences in simulation frameworks to achieve ecological validity of study outcomes and implications. Public repository for reproducibility: https://github.com/rUngruh/user_centered_choice_models.

2 Experimental Framework

Here, we discuss the experimental framework for our simulation study focused on choice models. (For analyses with additional metrics showing similar trends, see our repository.)

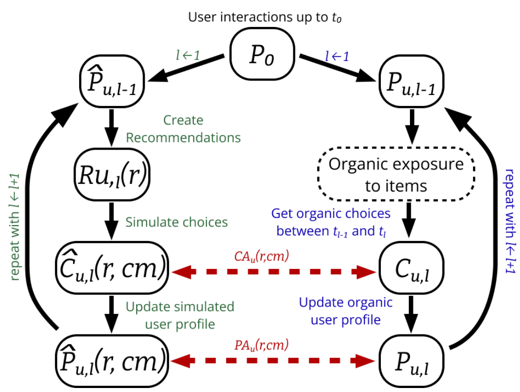

Figure 1: Simulation pipeline for each user u.

Simulation Pipeline. We iteratively simulate the recommendation and choice selection process for various combinations of RAs r and choice models cm. Each iteration l represents a timeframe of d days between two timestamps t_{l-1} and t_l.

The simulation pipeline (Fig. 1) for a given r–cm pair and a given dataset D consisting of user-item interactions for a set of users U follows established simulation setups [10, 36, 37, 44]:

(i) Initialize the user profiles ($\hat{P}_{u,0}$) for each user $u \in U$ with the list of rated items up to t_0. Set $l = 1$.
(ii) Create a rating matrix M_{l-1} based on $\hat{P}_{u,l-1}$ of all $u \in U$.
(iii) Train r using M_{l-1} and create recommendations ($R_{u,l}(r)$).
(iv) Simulate choices $\hat{C}_{u,l}(r, cm)$ using cm.
(v) Update the simulated user profiles ($\hat{P}_l = \hat{P}_{u,l-1} \cup \hat{C}_{u,l}$).
(vi) Repeat step 2–5 with $l \leftarrow l + 1$.

To compare the simulated choices with organic interaction patterns, we gather users' organic interactions for each timeframe: We extract organic choices $C_{u,l}$ from D for each iteration l (all items u consumed between two timestamps t_{l-1} and t_l in the dataset) and update the organic profile $P_{u,l}(r, cm)$ accordingly.

Dataset. We use **ML-20m** [23] due to its sequential structure [33]. We only consider ratings of movies with annotated genres. Each item is represented by a vector of genre weights; genres belonging to the item have a uniform weight distribution, others 0. We binarize ratings treating those higher than 3 as positive signals.

For our exploration, we consider user-item interactions from 2008 to 2010 as this 3-year window has the highest number of users who provided ratings throughout each year (2,461 users, 13,814 movies, 360,157 ratings). We set the start of the simulation t_0 = January 1st, 2009, resulting in one year for training and two for the simulation with 2.09 ratings on average per user per 30-day period after t_0. We set $d = 30$ days so that each simulation iteration represents a user receiving recommendations and consuming items every 30 days. Thus, 24 iterations capture the considered period.

RAs and Choice Models. We probe different RAs: **`Random`** and **`MostPop`** as non-personalized baselines; **`EASEr`** [43] since it commonly performs well across datasets [3]; **`RP`**$^3\beta$ [39] due to its high performance compared to popular RAs while being fast [3, 17]—a valuable property since training for multiple iterations is necessary; and **`ItemKNN`** [40] as a commonly used and fast baseline [3]. For deployment, we use the *Elliot* recommendation framework [4].

For universal choice models, we turn to prominent examples in recent RS simulation studies. Those include:

- `accept`$_c$: Chooses the top c recommendations [19, 21, 26, 44]. We set $c = 2$ to ensure a similar number of choices per iteration as observed in the organic behavior.
- `random`$_c$: Randomly selects c items [19, 30, 44]. We set $c = 2$.
- `ranked`$_\alpha$: Selects items based on ranked probabilities [10, 34, 37], where the probability of selecting an item i from $R_{u,l}(r)$ is $prob(i|R_{u,l}(r)) = e^{-\alpha * rank_t}$. This approach mimics the position bias [11, 31], $\alpha \in [0.6, 0.8, 1.0, 1.2, 1.4]$.
- `rankDCG`$_\alpha$: Picks items with ranked probabilities in line with DCG scores: $prob(i|R_{u,l}(r)) = \alpha \frac{1}{log_2(rank_i+1)}$, $\alpha \in [0.1, 0.2, 0.3, 0.5, 0.75]$.
- `popularity`$_c$: Picks the c most popular items from the recommendations, with popularity determined by the frequency items were consumed up to timestamp t_0 across all users [44] ($c = 2$).
- `RandomBrowsing`: As a variation to the browsing model from [27], this model selects the first item at random and iteratively picks additional items, with a logarithmically decreasing probability. Upon encountering a liked item—determined from organic interactions after t_0—the probability resets. Consecutive non-liked selections reduce the likelihood of continued exploration.

Metrics. To juxtapose simulated and organic choices using:

- Choice Alignment $CA_u(r, cm)$: Determines the degree to which simulated and natural item consumption overlap, based on the proportion of the 24 iterations in which at least one item in $\hat{C}_{u,l}(r, cm)$ can be found in $C_{u,l}$ for the same iteration. $CA_u(r, cm)$ emulates the *HIT* metric, which detects whether at least one true positive is predicted [45]—a facet that we probe at each iteration.
- Preference Alignment $PA_u(r, cm)$: Measures the degree to which preferences captured in $\hat{P}_{u,24}(r, cm)$ (resulting from the simulation across all 24 iterations) align with those in $P_{u,24}$ (constructed from organic consumption). We treat preferences as the genre distribution of items in a profile, i.e., the average genre vector for all items in the profile. Alignment between two vectors is measured as the complement of the Jensen-Shannon Divergence (JSD) [35]; higher values correspond to better alignment [cf. 47]. This metric resembles calibration, reflecting how closely genres of simulated consumption match those of actual consumption [34, 42].

Setup. We use the last month before t_0 as a validation set and conduct hyperparameter tuning for each r using Tree Parzen Estimator [8] for 20 iterations, utilizing parameter ranges as in [3]. Selecting the best parameters, we train each r using the entire data before t_0 and run the simulation pipeline for each r–cm pair. For each iteration, r is trained with the latest user profile, and $k = 10$ recommendations are created. In line with related studies [10, 44, 48], previously consumed items are not recommended.

3 Results & Discussion

Here, we present the results of our simulation study.

Monolithic Choice Models. To assess whether a universal choice model can reflect organic consumption patterns, we evaluate the alignment of simulated choices with real user consumption. Specifically, we report average $CA_u(r, cm)$ and $PA_u(r, cm)$ scores across all $u \in U$ for a given r–cm pair. The results in Table 1 show that there are choice models that lead to choices that significantly more closely align with organic consumption than others for a given RA (paired t-tests; $p < .05$). In terms of $CA_u(r, cm)$, $\texttt{rankDCG}_{0.75}$ aligns more closely with natural consumption than any other choice model for 3 out of 5 RAs. $CA_u(r, cm)$ scores are overall quite low, which we attribute to the low chance of having a matching item in $\hat{C}_{u,l}(r, cm)$ and $C_{u,l}$ at the same iteration, as both often only include ~ 2 items (out of an item corpus of 13,814 items). Based on $PA_u(r, cm)$, $\texttt{ranked}_{1.4}$ emerges as the best model across RAs.

Our results show that regardless of the RA, different choice models lead to different outcomes. This aligns with previous findings [24, 26], which highlight the influence of choice models on the characteristics of simulated choices. However, our analysis goes further by demonstrating that the specific choice model not only shapes simulated interactions but also directly impacts how well those simulated choices align with actual user consumption patterns.

These results reflect *average* effects, i.e., the impact of a choice model on the overall user base; potentially overlooking individuals. To examine whether outcomes for some users are not well represented by these trends, we analyze the alignment of universal choice models with *individual* user behavior. Specifically, we report the proportion of users for whom a given choice model provides the best or worst alignment with organic consumption, as measured by $PA_u(r, cm)$ for each RA. The results in Table 2 indicate that, regardless of the RA, $\texttt{ranked}_{1.4}$ most frequently provides the best alignment (also noted in Table 1); $\texttt{rankDCG}_{0.75}$ aligns most often the least. Although these choice models stand out, their dominance does not overshadow others: $\texttt{ranked}_{1.4}$ is the best-fitting model for only 25.8% of users when using ItemKNN, meaning that other models better represent behavior for the vast majority of the users.

To intuitively showcase that choice models do not equitably model all users, we depict in Fig. 2 $PA_u(r, cm)$ scores on a user level for ItemKNN (chosen as a simple and personalized baseline for context), computed for a sample of choice models that highlights broader effects, ranging from common approaches to those with higher $PA_u(ItemKNN, cm)$ scores (Table 1). $\texttt{ranked}_{1.4}$ stands out for most users by most closely aligning with their organic genre consumption. Still, on the right side of the graph—where $PA_u(\texttt{ItemKNN}, \texttt{ranked}_{1.4})$ scores are lower—for a large number of users, other choice models fare better. This suggests that the seemingly 'best' model does not capture behavior and preferences consistently for all users. Instead, several would be better represented if other choice models were used to simulate their choice behavior. For instance, consider $\texttt{accept}_2$, one of the most common choice models [19, 21, 26, 44]. Although it leads to less alignment for many users in comparison to $\texttt{ranked}_{1.4}$, Fig. 2 reveals a visibly prominent number of users for whom this model produces better alignment.

Table 1: Average $CA_u(r, cm)/PA_u(r, cm)$ across all $u \in U$. For each r, the best universal cm is bolded if scores are significantly higher than any other cm (paired t-test, $p < .05$). The mosaic model's values are reported and bolded if the metric scores differ significantly from any universal cm.

	Random	MostPop	ItemKNN	RP³β	EASEr
$\texttt{accept}_2$	0.001/0.840	0.028/0.853	0.025/0.852	0.029/0.859	0.027/0.858
$\texttt{random}_2$	0.001/0.840	0.026/0.849	0.027/0.850	0.027/0.858	0.028/0.852
$\texttt{popularity}_2$	**0.004**/0.853	0.028/0.846	0.025/0.845	0.028/0.847	0.029/0.846
randomBrowsing	0.001/0.838	0.027/0.846	0.024/0.849	0.028/0.856	0.029/0.852
$\texttt{rankDCG}_{0.1}$	0.001/0.862	0.020/0.868	0.019/0.870	0.020/0.873	0.022/0.872
$\texttt{rankDCG}_{0.2}$	0.001/0.854	0.020/0.862	0.022/0.864	0.026/0.870	0.025/0.865
$\texttt{rankDCG}_{0.3}$	0.001/0.848	0.023/0.858	0.023/0.857	0.026/0.863	0.027/0.858
$\texttt{rankDCG}_{0.5}$	0.002/0.838	0.032/0.844	0.029/0.848	0.034/0.855	0.033/0.851
$\texttt{rankDCG}_{0.75}$	0.002/0.829	**0.037**/0.830	**0.034**/0.842	0.038/0.852	**0.042**/0.843
$\texttt{ranked}_{0.6}$	0.001/0.849	0.022/0.861	0.023/0.859	0.024/0.866	0.025/0.861
$\texttt{ranked}_{0.8}$	0.001/0.855	0.022/0.862	0.020/0.866	0.021/0.870	0.024/0.865
$\texttt{ranked}_{1.0}$	0.001/0.858	0.022/0.861	0.021/0.869	0.022/0.871	0.025/0.868
$\texttt{ranked}_{1.2}$	0.002/0.862	0.018/0.866	0.021/0.870	0.019/0.872	0.023/0.871
$\texttt{ranked}_{1.4}$	0.001/**0.864**	0.023/**0.868**	0.016/**0.871**	0.024/**0.873**	0.021/**0.873**
mosaic	0.001/0.863	0.027/**0.870**	—	0.021/**0.878**	0.026/**0.875**

Table 2: Proportion (%) of users for whom each choice model provides the best/worst alignment according to $PA_u(r, cm)$ per RA. The most frequently best model for each RA is bolded.

	Random	MostPop	ItemKNN	RP³β	EASEr
$\texttt{accept}_2$	2.6/7.8	5.1/0.8	3.2/1.5	3.5/3.5	10.3/6.6
$\texttt{popularity}_2$	9.5/3.4	0.8/2.8	5.1/22.8	2.9/30.6	2.2/17.3
$\texttt{random}_2$	2.6/8.0	2.8/1.1	2.2/4.1	4.3/2.6	2.0/3.7
randomBrowsing	2.9/9.0	2.3/2.3	2.6/3.7	1.8/6.0	2.5/5.7
$\texttt{rankDCG}_{0.1}$	14.1/3.5	22.8/4.6	17.1/4.6	19.5/4.9	19.1/4.5
$\texttt{rankDCG}_{0.2}$	6.6/2.6	5.7/0.9	4.9/1.3	5.6/1.0	4.6/1.3
$\texttt{rankDCG}_{0.3}$	3.7/3.5	4.4/0.6	2.6/1.7	2.3/0.8	3.2/2.3
$\texttt{rankDCG}_{0.5}$	2.4/11.5	1.4/2.5	2.3/4.2	1.6/4.7	2.6/3.7
$\texttt{rankDCG}_{0.75}$	1.9/33.2	2.6/69.7	6.0/43.3	10.5/33.2	3.9/40.9
$\texttt{ranked}_{0.6}$	3.9/3.5	5.4/0.4	3.2/0.9	3.0/0.5	2.4/1.4
$\texttt{ranked}_{0.8}$	6.8/2.8	4.9/2.1	4.3/0.6	5.0/0.6	3.0/1.3
$\texttt{ranked}_{1.0}$	9.1/3.3	4.0/5.2	8.2/2.0	6.3/1.7	5.5/3.1
$\texttt{ranked}_{1.2}$	14.5/3.5	12.5/2.8	12.6/3.6	10.8/4.1	13.0/2.9
$\texttt{ranked}_{1.4}$	**19.3**/4.2	**25.2**/4.5	**25.8**/5.7	**22.9**/5.8	**25.6**/5.2

Mosaic Choice Models. Adopting a universal choice model for simulation leads to modeling that fails to accurately capture the behavior of a non-negligible number of users. This calls for holistic choice models that adapt to individual behavior. To explore if such a model improves alignment with natural consumption, we create the mosaic model, a holistic choice model based on insights from the previously discussed universal choice models. This aims to assume choice behavior that aligns 'best' with each user's natural consumption patterns. For this, we designate ItemKNN as the *baseline* RA due to its simplicity and personalized nature. We analyze which universal choice model resulted in the highest $PA_u(ItemKNN, cm)$ score for each user u. mosaic simulates this choice behavior for u at each iteration. We compare alignment scores, $PA_u(r, mosaic)$ and $CA_u(r, mosaic)$, resulting from a RS simulation with mosaic, with the respective scores for universal choice models for each RA (excluding ItemKNN, as it was used as a baseline).

Average $PA_u(r, cm)$ scores for mosaic in Table 1 are higher than those for any other universal choice model across RAs (except Random). Consistent with our results for universal choice models—where models tend to excel either on $PA_u(r, cm)$ or $CA_u(r, cm)$,

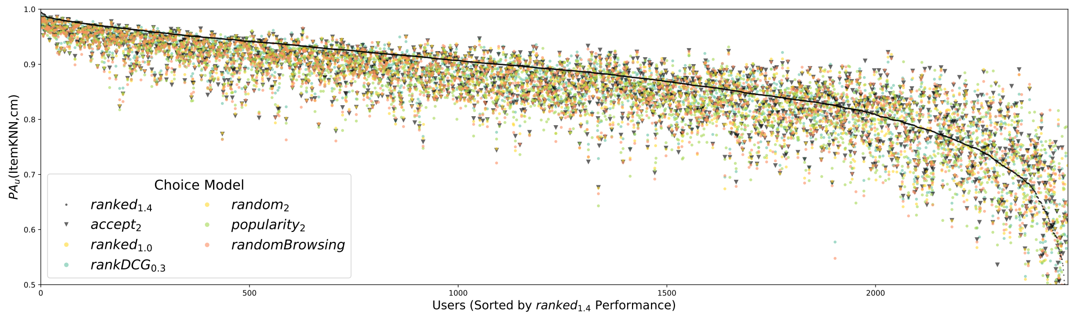

Figure 2: $PA_u(ItemKNN, cm)$ for users $u \in U$ and different cm.

but not both—$CA_u(r, cm)$ scores for `mosaic` remain comparable to those of universal models. These results reveal the limits of universal choice models in capturing individual consumption behaviors and show the potential of holistic modeling. By selecting the best-fitting choice behavior per user, `mosaic` achieves closer alignment with natural consumption patterns than any universal model.

Informed by reported outcomes to address our RQ, we conclude that a universal choice model cannot capture the full diversity of the user base; a holistic model, such as our `mosaic`, leads to choices that more closely resemble those of a non-homogeneous user base.

Implications. To conduct simulations of realistic user interactions with RS—and, consequently, to derive more meaningful insights—it is key to utilize choice models that accurately reflect the complexity and heterogeneity of real-world users. Our findings demonstrate that accounting for differences between users is essential. Even a naive modeling approach—`mosaic`—to creating holistic models based on simple choice models leads to better alignments than using the same model for all users. This underscores the need to further explore holistic choice models grounded in realistic user insights to improve alignment with users' organic consumption patterns further. Connecting analytics of past consumption patterns (i.e., training sets) with a nuanced understanding of behavioral patterns can enable simulation studies to reflect user–RS interactions closely.

Constructing a choice model that captures diverse and realistic consumption behaviors is challenging. We examined different measures of alignment, revealing that choice models vary in their ability to model user behavior. Even for universal choice models, the degree of alignment depends on the metric used; improving one measure may come at the expense of another. Thus, before attempting to enhance realism in a simulation study, the aspect of user behavior intended to be realistically simulated must be defined. To inform choice models, Chaney [9] suggests using *standard* choice models such as those utilized in economics and marketing [41]; we, instead, argue for the need to create choice models grounded on actual users, e.g., generated by data-driven approaches [25, 29] or based on psychological insights [2, 28].

Simulations are a powerful tool for assessing the long-term impact of RS, especially in data-limited settings or when real user access is limited [29]. As many studies focus on user impact [6, 16, 37], we argue that accounting for individual differences leads to more robust insights. Certain groups of users already face unequal treatment by RS [1, 15, 22, 38, 46]. Simulations have the potential to highlight these disparities [20, 37]; however, overlooking user diversity risks reinforcing inequalities rather than addressing them.

Limitations & Future Work. Our experimental setup follows standard practices. While we explicitly address the limitations of a 'general' view instead of a user-centered one, some broader limitations of RS simulation studies remain. The focus on the consumption of recommendations enables us to directly investigate the impact of choice models, but it overlooks naturally consumed items [9]. Further, utilized metrics capture different aspects of alignment and are affected in varying degrees by different choice models, highlighting their quality in detecting differences between users. However, they are influenced by profile sizes and the number of choices per iteration. For example, smaller initial profiles are more easily affected by new items in later iterations, leading to greater deviations. Additionally, the number of items picked by a choice model affects how much change can be captured. Furthermore, given the limited scope of this exploration, we focused on long-term effects and general developments across all simulated iterations. Further analysis requires more detailed explorations of changes in user behavior between iterations, as well as explorations of different datasets.

4 Conclusion

We compare alignment between RS simulations and actual behavior, probing how universal choice models capture organic consumption patterns. By analyzing individual users rather than overall trends, we reveal nuanced differences that may affect the realism of simulation outcomes and generalizability across users. As simulations in RS research become increasingly prominent, our work makes an argument for the development of realistic choice models, following efforts from studies of consumer behavior in market research aiming to move beyond "standard models" based on behavioral decision theories [12], but instead acknowledge the complexities of human psychology [7]. Our results show that in RS simulations, universal choice models do not reflect the intricacies of a diverse user base either, underscoring the need for holistic choice models that piece together user preferences like a *mosaic* of individual behaviors rather than forcing them into a monolithic mold.

Acknowledgments

Work supported by Grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and "ERDF, a way of making Europe."

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems (RecSys 2019), Available at https://doi.org/10.48550/arXiv.1907.13286* (2019).

[2] Icek Ajzen. 1998. Models of human social behavior and their application to health psychology. *Psychology and health* 13, 4 (1998), 735–739.

[3] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-n recommendation algorithms: A quest for the state-of-the-art. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 121–131.

[4] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2405–2414.

[5] Joeran Beel, Stefan Langer, Andreas Nürnberger, and Marcel Genzmehr. 2013. The impact of demographics (age and gender) and other user-characteristics on evaluating recommender systems. In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3*. Springer, 396–400.

[6] Alejandro Bellogín and Yashar Deldjoo. 2021. Simulations for novel problems in recommendation: analyzing misinformation and data characteristics. *SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research, in conjunction with 15th ACM Conference on Recommender Systems (RecSys 2021), Available at https://doi.org/10.48550/arXiv.2110.04037* (2021).

[7] Moshe Ben-Akiva, Daniel McFadden, Tommy Gärling, Dinesh Gopinath, Joan Walker, Denis Bolduc, Axel Börsch-Supan, Philippe Delquié, Oleg Larichev, Taka Morikawa, et al. 1999. Extended framework for modeling choice behavior. *Marketing letters* 10 (1999), 187–203.

[8] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* 24 (2011).

[9] Allison JB Chaney. 2021. Recommendation system simulations: A discussion of two key challenges. *SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research, in conjunction with 15th ACM Conference on Recommender Systems (RecSys 2021), Available at https://doi.org/10.48550/arXiv.2109.02475* (2021).

[10] Allison JB 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*. 224–232.

[11] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. A study of position bias in digital library recommender systems. *arXiv preprint arXiv:1802.06565* (2018).

[12] Hillel J Einhorn and Robin M Hogarth. 1981. Behavioral decision theory: Processes of judgement and choice. *Annual review of psychology* 32, 1981 (1981), 53–88.

[13] Michael D Ekstrand. 2021. Multiversal Simulacra: Understanding Hypotheticals and Possible Worlds Through Simulation. *SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research, in conjunction with 15th ACM Conference on Recommender Systems (RecSys 2021), Available at https://doi.org/10.48550/arXiv.2110.00811* (2021).

[14] Michael D. Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. 2021. SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, Humberto Jesús Corona Pampín, Martha A. Larson, Martijn C. Willemsen, Joseph A. Konstan, Julian J. McAuley, Jean Garcia-Gathright, Bouke Huurnink, and Even Oldridge (Eds.). ACM, 803–805. doi:10.1145/3460231.3470938

[15] 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 *Conference on fairness, accountability and transparency*. PMLR, 172–186.

[16] Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. 2022. Exposure inequality in people recommender systems: the long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 194–204.

[17] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

[18] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jason Yoon. 2019. Artist and style exposure bias in collaborative filtering based music recommendations. *Workshop on Designing Human-Centric MIR Systems, in conjunction with 20th International Society for Music Information Retrieval Conference (ISMIR 2019), Available at: https://doi.org/10.48550/arXiv.1911.04827* (2019).

[19] Andres Ferraro, Michael D Ekstrand, and Christine Bauer. 2024. It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 884–889.

[20] Andres Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring longitudinal effects of session-based recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 474–479.

[21] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 249–254.

[22] Mustansar Ali Ghazanfar and Adam Prügel-Bennett. 2014. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications* 41, 7 (2014), 3261–3275.

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

[24] Naieme Hazrati and Francesco Ricci. 2022. Recommender systems effect on the evolution of users' choices distribution. *Information Processing & Management* 59, 1 (2022), 102766.

[25] Naieme Hazrati and Francesco Ricci. 2022. Simulating users' interactions with recommender systems. In *Adjunct proceedings of the 30th acm conference on user modeling, adaptation and personalization*. 95–98.

[26] Naieme Hazrati and Francesco Ricci. 2024. Choice models and recommender systems effects on users' choices. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 109–145.

[27] Katja Hofmann, Anne Schuth, Alejandro Bellogin, and Maarten De Rijke. 2014. Effects of position bias on click-based recommender evaluation. In *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36*. Springer, 624–630.

[28] Christoffer Holmgård, Antonios Liapis, Julian Togelius, and Georgios N Yannakakis. 2014. Evolving personas for player decision modeling. In *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 1–8.

[29] Chih-Wei Hsu, Martin Mladenov, Ofer Meshi, James Pine, Hubert Pham, Shane Li, Xujian Liang, Anton Polishko, Li Yang, Ben Scheetz, et al. 2024. Minimizing live experiments in recommender systems: User simulation to evaluate preference elicitation policies. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2925–2929.

[30] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015), 427–491.

[31] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[32] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2018. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers* 20 (2018), 1241–1265.

[33] Anton Klenitskiy, Anna Volodkevich, Anton Pembek, and Alexey Vasilev. 2024. Does It Look Sequential? An Analysis of Datasets for Evaluation of Sequential Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1067–1072.

[34] Oleg Lesota, Jonas Geiger, Max Walder, Dominik Kowald, and Markus Schedl. 2024. Oh, Behave! Country Representation Dynamics Created by Feedback Loops in Music Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1022–1027.

[35] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151.

[36] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. 2021. T-RECS: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959* (2021).

[37] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[38] 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. In *The thirty-third international flairs conference*.

[39] Bibek Paudel, Fabian Christoffel, Chris Newell, and Abraham Bernstein. 2016. Updatable, accurate, diverse, and scalable recommendations for interactive applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–34.

[40] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[41] H Simon. 1957. A Behavioral Model of Rational Choice. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social*

Setting/Wiley (1957).

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

[43] Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*. 3251–3257.

[44] Zoltán Szlávik, Wojtek Kowalczyk, and Martijn Schut. 2011. Diversity measurement of recommender systems under different user choice models. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 369–376.

[45] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. 2021. Quality metrics in recommender systems: Do we calculate metrics consistently?. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 708–713.

[46] Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. The Impact of Mainstream-Driven Algorithms on Recommendations for Children. In *European Conference on Information Retrieval*. Springer, 67–84.

[47] Saúl Vargas and Pablo Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender systems*. 145–152.

[48] Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. 2021. Measuring recommender system effects with simulated users. *Second Workshop on Fairness, Accountability, Transparency, Ethics and Society on the Web (FATES 2020), Available at: https://doi.org/10.48550/arXiv.2101.04526* (2021).

[49] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101.