Measuring Item Fairness in Next Basket Recommendation: A Reproducibility Study

Yuanna Liu[1](✉), Ming Li[2], Mozhdeh Ariannezhad[3], Masoud Mansoury[1,4], Mohammad Aliannejadi[1], and Maarten de Rijke[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
{y.liu8,m.mansoury,m.aliannejadi,m.derijke}@uva.nl
[2] AIRLab, Amsterdam, The Netherlands
m.li@uva.nl
[3] Booking.com, Amsterdam, The Netherlands
mozhdeh.ariannezhad@booking.com
[4] Discovery Lab, Elsevier, Amsterdam, The Netherlands

Abstract. Item fairness of recommender systems aims to evaluate whether items receive a fair share of exposure according to different definitions of fairness. Raj and Ekstrand [26] study multiple fairness metrics under a common evaluation framework and test their sensitivity with respect to various configurations. They find that fairness metrics show varying degrees of sensitivity towards position weighting models and parameter settings under different information access systems. Although their study considers various domains and datasets, their findings do not necessarily generalize to next basket recommendation (NBR) where users exhibit a more repeat-oriented behavior compared to other recommendation domains. This paper investigates fairness metrics in the NBR domain under a unified experimental setup. Specifically, we directly evaluate the item fairness of various NBR methods. These fairness metrics rank NBR methods in different orders, while most of the metrics agree that repeat-biased methods are fairer than explore-biased ones. Furthermore, we study the effect of unique characteristics of the NBR task on the sensitivity of the metrics, including the basket size, position weighting models, and user repeat behavior. Unlike the findings in [26], Inequity of Amortized Attention (IAA) is the most sensitive metric, as observed in multiple experiments. Our experiments lead to novel findings in the field of NBR and fairness. We find that Expected Exposure Loss (EEL) and Expected Exposure Disparity (EED) are the most robust and adaptable fairness metrics to be used in the NBR domain.

1 Introduction

Fairness of information access systems is increasingly drawing attention [8]. Such systems are not only required to have high accuracy, but they should also be fair to both users and providers. Usually, items are exposed to users as ranked lists based on a *relevance* or *utility* score. The *exposure* of items influence users' browsing, clicking, and

M. Ariannezhad—Work done when the author was a member of AIRLab at the University of Amsterdam.

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14611, pp. 210–225, 2024.
https://doi.org/10.1007/978-3-031-56066-8_18

purchase behavior [26]. However, items may not receive fair exposure due to algorithmic or data biases [22]. To understand different notions and aspects of fairness, many fairness definitions and metrics have been proposed. Raj and Ekstrand [26] unify several fair ranking metrics under a common evaluation framework and compare them empirically using different information access tasks, viz. book recommendation and scholarly article retrieval and re-ranking. The authors focus on provider-side group fairness of ranked lists and design a sensitivity analysis to evaluate the robustness of fairness metrics w.r.t. position weighting models and parameter settings. Importantly, they find that *fairness metrics show different patterns of sensitivity for different search and recommendation tasks* — this means that the lessons learned in one search and recommendation scenario usually cannot be directly and completely transferred to another domain.

Item Fairness and Next Basket Recommendation. Recently, there has been increased interest in the task of next basket recommendation (NBR) [3]. NBR models users' preferences based on a sequence of historical sets of items (baskets, playlists, reading lists, ...) and then predicts, for each user, a set of items they are likely to purchase, listen to, or read next. The NBR task is important because it is relevant across a broad range of domains, from e-commerce and travel to education and entertainment. We are interested in item fairness in the context of NBR: is the exposure assigned by an NBR system fairly distributed and do the main findings from [26] generalize to the NBR task? In particular, we focus on the following findings from [26] and examine their validity in the context of NBR:

– Metrics usually disagree on the orderings of methods; we expect to find similar patterns for NBR baselines.
– For the FairTREC full retrieval task, changing the size of the ranked list has no impact on the performance of fairness measurements. But the fairness metrics Attention-Weighted Rank Fairness (AWRF) [27], IAA, EEL, EED, and Expected Exposure Relevance (EER) are stable on the GoodReads recommendation task, while Demographic Parity (logDP), Exposed Utility Ratio (logEUR), and FAIR [34] change with different-sized ranked lists and even alter the relative order of recommendation algorithms. We change the basket size and focus on the magnitude of the change of metric values and whether the fairness order of NBR methods will be affected.
– When used with different position weighting models, fairness metrics show different degrees of robustness. Position weighting models influence metric values and algorithms' ordering, especially for EEL and Realized Utility Ratio (logRUR). For the NBR scenario, we experiment with the document-based click-through rate model (DCTR) [5], which means each position has the same exposure in a list. We expect to identify the most unstable metrics.

Item Fairness and Repeat Behavior. Going beyond the above generalizability questions, there is a specific characteristic of NBR that may affect item fairness in unknown ways: in many NBR scenarios, users display a significant amount of repeat behavior, whereby they purchase or consume the same item multiple times. Some users mostly purchase repeat items (i.e., items that they consumed before) as part of their next basket, while others tend to explore items (items that they never bought before). Ekstrand et al. [7] suggest we need to look at metric values across different user groups rather than at the population as a whole. Following this advice, we design experiments to assess item fairness across different user groups with varying degrees of repeat purchase behavior.

To the best of our knowledge, we are the first to study fairness in next basket recommendation. We focus on four research questions:

(RQ1) What is the fairness ranking of NBR methods? Are the lessons observed in [26] valid for the NBR task?
(RQ2) Which fairness metrics are more robust to basket size?
(RQ3) Which fairness metrics are more robust to position weighting models for NBR?
(RQ4) How does repeat purchase behavior affect item fairness?

2 Related Work

Next Basket Recommendation. NBR systems make personalized product recommendations to users based on their historical baskets. Repeat purchase behavior is a prominent pattern of NBR, which makes it different from typical recommendation domains, such as movie recommendation [11]. Since the models designed for recommending repeat items and explore items vary significantly, recent work proposes more targeted task settings and algorithms [14, 17, 18] to address the unique challenges in this domain by modeling the repeat behavior of the users while trying to optimize the explore behavior of the users.

A significant amount of research concentrates on employing neural networks to learn representations for sequences of baskets. The effectiveness of recurrent neural networks (RNNs) in sequential modeling has led to their application in NBR. Yu et al. [32] propose a dynamic recurrent basket model, feeding a series of basket representations to the recurrent architecture to obtain the dynamic representation of a user. Le et al. [16] model correlation information to augment the representation of basket sequence. Yu et al. [33] propose a model based on graph neural networks (GNNs) for temporal sets prediction, where sets are constructed as weighted graphs and a graph convolutional network is applied to capture relationships among elements in each set. Ariannezhad et al. [2] analyze users' repeat consumption behavior and propose a repeat consumption-aware neural network for NBR.

However, recent nearest neighbor-based methods show more effective performance and efficiency than neural network-based baselines on NBR. Faggioli et al. [9] propose recency-aware, user-wise popularity and incorporate it into both user- and item-based collaborative approaches. Hu et al. [13] integrate temporal dynamics into personalized item frequency and then use a user-KNN method to make predictions. Naumov et al. [24] improve on [13] by considering time intervals between interactions.

Li et al. [19] reproduce NBR methods in a unified experimental setting and propose a new angle to evaluate the performance obtained from repeat items and explore items, respectively. According to whether a method tends to recommend repeat items or explore items, it is called repeat-biased or explore-biased. Our work reproduces several representative NBR methods in a unified experimental setting and focuses on measuring item fairness and exploring the impact of repeat behavior on fairness.

Item Fairness in Recommender Systems. Fairness research raises challenges for information access systems characterized by (i) a multi-stakeholder nature, (ii) a rank-based problem setting, (iii) the requirement of personalization in many cases, and (iv) the role of user feedback [8]. Wang et al. [30] collect fairness definitions in the recommendation literature and provide views of classifying fairness issues. Consumer fairness cares about whether users receive comparable recommendation quality from the system [21]. In contrast, provider fairness focuses on how to assign reasonable exposure to each document, content provider, or group [26]. The allocation criteria can be with reference to a distribution [27,31,34] or proportional to merit [6].

Some metrics have primarily undergone testing using small and/or synthetic datasets, and they encounter challenges in handling complex real-world information access applications where incomplete data and extreme cases happen [26]. Raj and Ekstrand [26] implement experiments on book recommendation and scholarly article retrieval tasks and test the sensitivity of fairness metrics towards parameter setting. Kowald et al. [15] reproduce the analyses of [1] to investigate how popularity bias causes unfairness for both long-tail items and low-mainstream users in the context of music recommendation. We can observe that the conclusions about the fairness evaluation in one domain cannot be completely transferred to another domain.

NBR has distinguishing characteristics, i.e., repeat items figure prominently amongst the recommended results and make a considerable contribution to accuracy [19]. There is no prior work studying fairness in the context of NBR. Li et al. [20] argues that the frequency bias harms the fairness of NBR system. To fill this gap, we reproduce the fairness evaluation and sensitivity experiments of [26] to (i) see if the patterns they found can be generalized to NBR domain, and (ii) select robust metrics suitable for NBR.

3 Reproducibility Setup

3.1 Problem Formulation

Our experiments concern two main parts: NBR and fairness evaluation. Firstly, in next basket recommendation, we denote D as the item set and Q as the user set. A basket is a set of items $B = [d_1, d_2, \ldots, d_m]$, where $d_i \in D$ denotes an item from the basket B. Items have no temporal order and hold equal significance in a basket. For each user $q \in Q$, there is a sequence of historical purchase baskets $B^q = [B_1^q, B_2^q, \ldots, B_n^q]$, where B_i^q indicates the i-th basket purchased by the user. The goal of NBR is to predict

Table 1. Notation used in the paper; adapted from [26].

$d \in D$	item	$q \in Q$	user
L	ranked list of N items (predicted basket)	$L^{-1}(i)$	the item in position i of basket L
$L(d)$	rank of item d in L	$y(d\|q)$	relevance of d to q
$\hat{y}(d\|q)$	predicted relevance of d to q	$G(L)$	group alignment matrix for items in L
G^+	popular group	G^-	unpopular group
$\mathbf{a}_L$	exposure vector for items in L	ϵ_L	the exposure of groups in L $(G(L)^T \mathbf{a}_L)$

the next basket L for each user. Then, we evaluate the item fairness of predicted baskets among all users using the fair ranking metrics in Sect. 3.4. Specifically, the predicted basket L can be seen as a ranked list where the ranking is based on the user-specific relevance score $\hat{y}(d \mid q)$ predicted by each NBR method. The goal of item fairness is to measure whether the exposure is fairly distributed among the groups according to a specific principle. The notation used in this paper is summarized in Table 1.

3.2 Datasets

Following [2,12,25], we use three publicly available datasets for our experiments: (i) Instacart,[1] which contains a sample of over three million grocery orders from users. Items belonging to the same order form a basket. (ii) Dunnhumby,[2] which includes household-level transactions over two years from 2,500 households. (iii) TaFeng,[3] which contains transaction data from a Chinese grocery store from November 2000 to February 2001. We treat all transactions of a user within a day as a basket.

For each dataset, we remove users with fewer than three baskets and items bought fewer than five times [2]. Due to the large number of baskets in Instacart, the calculation of some methods exceeds the memory, therefore we randomly sampled 20,000 users from Instacart before filtering [24]. Table 2 shows the statistics of three datasets after preprocessing. Avg. repeat ratio refers to the average proportion of repeat items in the ground truth baskets [19]. We split each dataset following [2,9,24]. The training baskets consist of all baskets of users except the last one. For users who have more than 50 baskets in the training data, we only consider their last 50 baskets in the training set [19]. The last baskets of all the users are split into 50% validation set and 50% test set.

3.3 NBR methods

We select the following representative NBR methods with open-source code, including frequency-based, nearest neighbor-based, and deep learning-based methods. According to the classification of NBR methods in [19], G-TopFreq and Dream are explore-biased methods (the recommended baskets are skewed towards explore items averaged over all users), which are also popularity-based methods. Other methods are repeated-biased

Table 2. Statistics of datasets after preprocessing.

Dataset	#Users	#Items	#Baskets	Avg. #baskets/user	Avg. #items/basket	Avg. repeat ratio
Instacart	19,210	29,399	305,582	15.91	10.06	0.60
Dunnhumby	2,482	37,162	107,152	43.17	10.07	0.43
TaFeng	10,182	15,024	82,387	8.09	6.14	0.21

[1] https://www.kaggle.com/c/instacart-market-basket-analysis/data.

[2] https://www.dunnhumby. com/source-files/.

[3] https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset.

Table 3. Summary of fairness metrics; adapted from [26].

Category	Metrics	Goal	Binomial	More Fair
Equal opportunity	logEUR	Exposure proportional to relevance	✓	○
	logRUR	Click-through rate proportional to relevance	✓	○
	IAA	Exposure proportional to predicted relevance	×	↓
	EEL,EER	Exposure matches ideal (from relevance)	×	EEL ↓, EER ↑
Statistical parity	EED	Exposure well-distributed	×	↓
	logDP	Exposure equal across groups	✓	○

methods (the recommended baskets are skewed towards repeat items). We cover various types of NBR method and study their fairness performance.

Frequency-Based Methods. (i) G-TopFreq recommends the most popular k items across all historical purchases to all users. (ii) P-TopFreq counts the k products with the highest frequency in each user's historical purchase records. (iii) GP-TopFreq firstly recommends personal history most popular items and then fills the empty slot with global most popular items.

Nearest Neighbor-Based Methods. (i) TIFUKNN [13] integrates temporal dynamics modeled by time-decayed weights to generate user representations. Then, the target user representation and its nearest neighbors are combined to make the prediction. (ii) UP-CF@r [9] incorporates recency-aware user-wise popularity in a collaborative filtering framework.

Deep Learning-Based Methods. (i) Dream [32] is an RNN-based model that learns a dynamic representation of a user and captures global sequential features among baskets. (ii) DNNTSP [33] constructs weighted graphs to learn basket-level element relationships. The attention mechanism is used to learn the temporal dependencies of sets and elements. Static and dynamic representations are fused by a gated updating mechanism. (iii) ReCANet [2] focuses on repeat consumption, combines user-item representations with historical consumption patterns, and models temporal signals by LSTM layers.

3.4 Fair Ranking Metrics

Raj and Ekstrand [26] unify several fair ranking metrics in a common framework and notation. They clarify the limitation of these metrics when applying to practical information access systems, and test the robustness of these metrics towards design and parameter choices. We follow the notation and fairness implementation of this paper. The fairness metrics shown in Table 3 are selected for the following reasons: (i) The predicted basket can be formulated as a ranking L. These metrics are well-known fairness metrics for rankings. (ii) These metrics cover two categories of fairness definitions: *statistical parity* (aimed at ensuring comparable exposure among groups) and *equal opportunity* (aimed at promoting equal treatment based on merit or utility, regardless of group membership) [26]. (iii) Since fair exposure is unlikely to be satisfied in any single ranking [4] and it is more practical to pursue fair exposure of items to overall users, we only consider the fairness of multiple rankings for NBR setting.

Assume $\pi(L \mid q)$ is a user-dependent distribution and $\rho(q)$ is a distribution over users, overall rankings among all the users follow the distribution of $\rho(q)\pi(L \mid q)$. $\epsilon_L = G(L)^T \mathbf{a}_L$ is the group exposure within a single ranking. Its expected value $\epsilon_\pi = E_{\pi\rho}[\epsilon_L]$ is the group exposure among all the rankings.

Equal Opportunity. Singh and Joachims [28] propose two ratio-based metrics. Exposed Utility Ratio (EUR) quantifies the deviation from the objective that the exposure of each group is proportional to its utility $Y(G)$:

$$\mathrm{EUR} = \frac{\epsilon_\pi\left(G^+\right)/Y\left(G^+\right)}{\epsilon_\pi\left(G^-\right)/Y\left(G^-\right)}. \tag{1}$$

Realized Utility Ratio (RUR) [28] models actual user engagement, the click-through rates for the groups $\Gamma(G)$ are proportional to their utility:

$$\mathrm{RUR} = \frac{\boldsymbol{\Gamma}\left(G^+\right)/Y\left(G^+\right)}{\boldsymbol{\Gamma}\left(G^-\right)/Y\left(G^-\right)}. \tag{2}$$

Biega et al. [4] propose Inequality of Amortized Attention (IAA), which takes the L_1 norm of the difference between cumulative exposure and cumulative system-predicted relevance $\hat{y}(d|q)$ for each group. For consistency, we normalize predicted relevance scores to be in the same range as exposure values:

$$\mathrm{IAA} = \|\epsilon_\pi - \hat{Y}\|_1. \tag{3}$$

Diaz et al. [6] define the target exposure ϵ^* as the expected exposure under the ideal policy. Expected Exposure Loss (EEL) is the distance between expected exposure and target exposure:

$$\mathrm{EEL} = \|\epsilon_\pi - \epsilon^*\|_2^2 = \|\epsilon_\pi\|_2^2 - 2{\epsilon_\pi}^T\epsilon^* + \|\epsilon^*\|_2^2. \tag{4}$$

EEL can be decomposed into EER $= 2{\epsilon_\pi}^T\epsilon^*$ (measuring the alignment of exposure and relevance) and Expected Exposure Disparity (EED).

Statistical Parity. EED [6] measures the inequality in exposure distribution across groups:

$$\mathrm{EED} = \|\epsilon_\pi\|_2^2. \tag{5}$$

Demographic Parity (DP) [28] measures the ratio of average exposure given to the two groups:

$$\mathrm{DP} = \epsilon_\pi\left(G^+\right)/\epsilon_\pi\left(G^-\right). \tag{6}$$

Following [26], DP is reformulated as logDP $= \log(\epsilon_\pi(G^+) + 10^{-6}) - \log(\epsilon_\pi(G^-) + 10^{-6})$ to address the empty-group problem and enhance interpretability. logEUR and logRUR are defined in the same way.

Table 4. Position weighting models for computing $\mathbf{a}_L$; adapted from [26].

Metric	Model	Formula	Parameters
IAA	Geometric	$\gamma(1-\gamma)^{L(d)-1}$	patience γ
logDP, logEUR, logRUR	Logarithmic	$1/\log_2 \max\{L(d), 2\}$	–
EER, EED,EEL	RBP [23]	$\gamma^{L(d)}$	patience γ
EER, EED,EEL	Cascade	$\gamma^{L(d)-1} \prod_{j \in [0,L(d))} [1-\phi(y(L^{-1}(j) \mid y))]$	patience γ stopping probability function ϕ
–	DCTR	$1/\lvert L(d)\rvert$	–

3.5 Position Weighting Models

Since users are likely to pay decreasing attention to lower-ranked items (position bias) [4], position weighting models are required when computing exposure [6]. These metrics explicitly represent the position weighting model as a position weight vector $\mathbf{a}_L$, as shown in Table 4 [26]. In NBR, many works [14,29,32] treat recommended results as ranked lists and evaluate NBR methods based on ranking metrics, e.g. normalized discounted cumulative gain (NDCG). Therefore, it is required to take the position of the recommended items into account when applying fair ranking metrics in NBR.

3.6 Implementation Details

In order to evaluate group fairness using the above metrics, following previous work [10], the division of items is based on item popularity, i.e., the number of purchases in the historical baskets of all users. We define the top 20% items with the highest purchase as popular group G^+ and the remaining 80% of the items as unpopular group G^-. The default basket size is set to 10. For all NBR baselines, we perform a grid search based on the hyperparameter ranges given in the original papers to find the optimal hyperparameters using the validation set. For TIFUKNN, the number of nearest neighbors k is tuned on $\{100, 300, 500, 900, 1100, 1300\}$, the number of groups m is chosen from $\{3, 7, 11, 15, 19, 23\}$, the within-basket time-decayed ratio r_b and the group time-decayed ratio r_g are selected from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, and the fusion weight α is tuned on $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. For UP-CF@r, recency window r is tuned on $\{1, 5, 10, 25, 100, \infty\}$, locality q is tuned on $[1, 5, 10, 50, 100, 1000]$, and asymmetry α is tuned on $\{0, 0.25, 0.5, 0.75, 1\}$. For Dream and DNNTSP, the item embedding size is tuned on $\{16, 32, 64, 128\}$. For ReCANet, user embedding size and item embedding size are tuned on $\{16, 32, 64, 128\}$. We run each method 5 times with 5 fixed random seeds to eliminate the random initialization effect and report the average results. Following [13,19,33], we use three accuracy metrics. Recall assesses the capacity to retrieve all relevant items. NDCG measures ranking quality, which considers sequence order by giving the lower-ranked items a discount. Personalized Hit Ratio (PHR) computes the proportion of predicted baskets that include at least one item from the ground truth basket. We release our code and hyperparameters at https://github.com/lynEcho/NBR-fairness.

4 Experiments and Results

Fairness Valuation. To answer **RQ1**, we measure item fairness of the recommendation results obtained by each NBR method on Instacart, Dunnhumby, and TaFeng. We

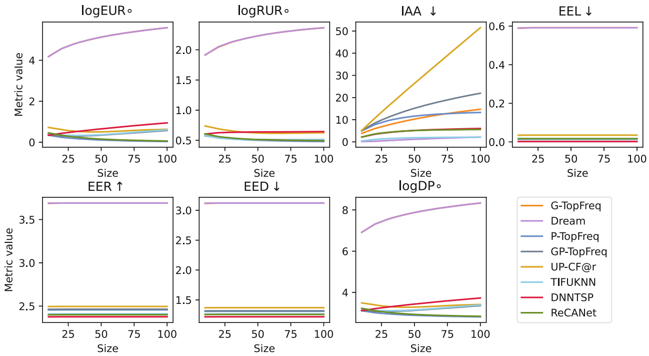


Fig. 1. Performance of different NBR methods in terms of item fairness with varying basket size. Each plot represents one item fairness metric.

implement the fairness metrics following the configurations from their original papers. Table 5 compares the NBR methods in terms of accuracy and fairness metrics. We make the following observation: (i) In most cases, repeat-biased methods exhibit more effectiveness in Recall, NDCG, and PHR, compared to explore-biased methods. (ii) Consistent with observations in [26], our experimental results lead to different orderings of the NBR methods when ranked using the fairness metrics. For instance, on Instacart, EEL indicates DNNTSP and P-TopFreq as the top two methods, whereas logRUR ranks TopFreq and GP-TopFreq as the top two. (iii) Metrics that consider equal opportunity, namely, logEUR, logRUR, and EEL deem repeat-biased methods fairer than the explore-biased ones (G-TopFreq and Dream). Surprisingly, according to EER, G-TopFreq and Dream are fairer than the repeat-biased methods. Both G-TopFreq and Dream only recommend popular items, which is unfair. (iv) Statistical-parity metrics, namely, EED and logDP, agree that repeat-biased methods are fairer than explore-biased methods. Because statistical parity aims at ensuring equal exposures across groups, however G-TopFreq and Dream only recommend popular items.

Basket Size. To answer **RQ2**, we change the basket size from 10 to 100 for each dataset. Figure 1 shows the variation of fairness values as basket size changes on Instacart.[4] It can be observed that: (i) IAA shows the highest sensitivity to basket size, as we see the highest increase in its value as the basket size grows. We relate it to the fact that as more items are considered in the list, it leads to more accumulation of gaps between position weights and predicted relevance. And the fairness order of P-TopFreq and G-TopFreq even changes. This differs from [26], where IAA exhibits stable behavior in the GoodReads recommendation task. (ii) For logDP, logEUR, and logRUR, we observe a mixed behavior for different basket sizes. As basket size increases, DNNTSP, GTop-Freq, and Dream become less fair in terms of all three metrics; PTop-Freq and ReCANet become fairer. GP-TopFreq, UF-CF@r, and TIFUKNN, however, exhibit

[4] We observe a similar trend on the Dunnhumby and TaFeng datasets. Because of space limitations, we only report the results on the Instacart dataset.

Table 5. Overall performance comparison of NBR methods.

Data-		Accuracy			Equal opportunity					Statistical parity	
set	Method	Recall↑	NDCG↑	PHR↑	logEUR○	logRUR○	IAA↓	EEL↓	EER↑	EED↓	logDP○
Instacart	G-TopFreq	0.0704	0.0817	0.4600	4.1803	1.9140	3.8532	0.5899	3.6884	3.1169	6.9148
	Dream	0.0704	0.0817	0.4600	4.1803	1.9143$_{(0.0001)}$	0.1466$_{(0.0053)}$	0.5895	3.6879	3.1160$_{(0.0001)}$	6.9148
	P-TopFreq	0.3143	0.3339	0.8447	0.3419	0.5756	4.9492	0.0138	2.4561	1.3086	3.1112
	GP-TopFreq	0.3150	0.3343	0.8460	0.3616	0.5764	5.1071	0.0144	2.4633	1.3164	3.1313
	UP-CF@r	0.3377	0.3582	0.8586	0.7217	0.7383	5.1405	0.0351	2.4951	1.3688	3.4917
	TIFUKNN	0.3456	0.3657	0.8639	0.4275	0.5800	0.4227	0.0154	2.4095	1.2635	3.1972
	DNNTSP	0.3295$_{(0.0002)}$	0.3434$_{(0.0005)}$	0.8581$_{(0.0001)}$	0.3458$_{(0.0171)}$	0.6050$_{(0.0104)}$	2.1511$_{(0.0462)}$	0.0020$_{(0.0004)}$	2.3760$_{(0.0070)}$	1.2166$_{(0.0075)}$	3.1154$_{(0.0171)}$
	ReCANet	0.3490$_{(0.0001)}$	0.3699$_{(0.0001)}$	0.8668$_{(0.0003)}$	0.4588$_{(0.0184)}$	0.6069$_{(0.0056)}$	2.1972$_{(0.0691)}$	0.0173$_{(0.0004)}$	2.4000$_{(0.0015)}$	1.2559$_{(0.0019)}$	3.2283$_{(0.0184)}$
Dunnhumby	G-TopFreq	0.0897	0.0798	0.3795	6.2326	3.5166	2.8829	1.0020	3.1503	3.2310	6.6211
	Dream	0.0896$_{(0.0003)}$	0.0759$_{(0.0002)}$	0.3873$_{(0.0013)}$	6.2326	3.5511$_{(0.0020)}$	0.3248$_{(0.0035)}$	0.9940	3.1427	3.2154	6.6211
	P-TopFreq	0.1628	0.1562	0.5399	0.9611	1.6879	4.6402	0.4039	2.7369	2.2195	3.2579
	GP-TopFreq	0.1628	0.1562	0.5399	0.9620	1.6879	4.6480	0.4039	2.7369	2.2196	3.2588
	UP-CF@r	0.1699	0.1639	0.5536	1.2771	1.6287	4.4460	0.4279	2.7297	2.2364	3.5630
	TIFUKNN	0.1763	0.1683	0.5729	1.6677	1.8252	0.9255	0.6236	2.8502	2.5525	3.9346
	DNNTSP	0.0871$_{(0.0019)}$	0.0792$_{(0.0016)}$	0.4303$_{(0.0048)}$	0.5513$_{(0.0268)}$	1.7653$_{(0.0506)}$	0.1037$_{(0.0363)}$	0.2545$_{(0.0126)}$	2.7432$_{(0.0130)}$	2.0765$_{(0.0255)}$	2.8581$_{(0.0263)}$
	ReCANet	0.1730$_{(0.0011)}$	0.1625$_{(0.0010)}$	0.5655$_{(0.0018)}$	1.0334$_{(0.0204)}$	1.6296$_{(0.0334)}$	0.4417$_{(0.0127)}$	0.4775$_{(0.0075)}$	2.7926$_{(0.0066)}$	2.3489$_{(0.0139)}$	3.3279$_{(0.0197)}$
TaFeng	G-TopFreq	0.0812	0.0893	0.2565	5.4916	2.4948	2.9267	1.0932	3.1283	3.3155	7.6939
	Dream	0.0888$_{(0.0032)}$	0.0924$_{(0.0011)}$	0.2798$_{(0.0059)}$	5.4916	2.5259$_{(0.0072)}$	0.6672$_{(0.0055)}$	1.0929$_{(0.0001)}$	3.1280$_{(0.0001)}$	3.3150$_{(0.0002)}$	7.6939
	P-TopFreq	0.1087	0.0983	0.3608	0.4059	0.8820	5.5137	0.2735	2.7269	2.0944	2.4192
	GP-TopFreq	0.1183	0.1018	0.3740	0.4357	0.8986	5.6991	0.2775	2.7335	2.1050	2.4492
	UP-CF@r	0.1406	0.1222	0.4325	1.6427	1.4053	4.7278	0.6896	2.9401	2.7237	3.6597
	TIFUKNN	0.1618	0.1419	0.4697	1.1420	1.3862	0.8630	0.6299	2.8835	2.6074	3.1573
	DNNTSP	0.1483$_{(0.0004)}$	0.1239$_{(0.0004)}$	0.4482$_{(0.0011)}$	1.4265$_{(0.0645)}$	1.5422$_{(0.0444)}$	0.3508$_{(0.0332)}$	0.4712$_{(0.0246)}$	2.8531$_{(0.0121)}$	2.4183$_{(0.0367)}$	3.4427$_{(0.0647)}$
	ReCANet	0.1287$_{(0.0002)}$	0.1188$_{(0.0002)}$	0.4195$_{(0.0005)}$	0.9450$_{(0.0045)}$	1.2147$_{(0.0080)}$	0.1644$_{(0.0098)}$	0.4343$_{(0.0072)}$	2.8093$_{(0.0042)}$	2.3376$_{(0.0114)}$	2.9597$_{(0.0045)}$

↑ indicates that larger value means better fairness; ↓ indicates that smaller value means better fairness; ○ means that the closer the value is to 0, the better the fairness. The number in each cell is mean, and the number in parentheses is standard deviation. The missing standard deviation represents 0.0000.

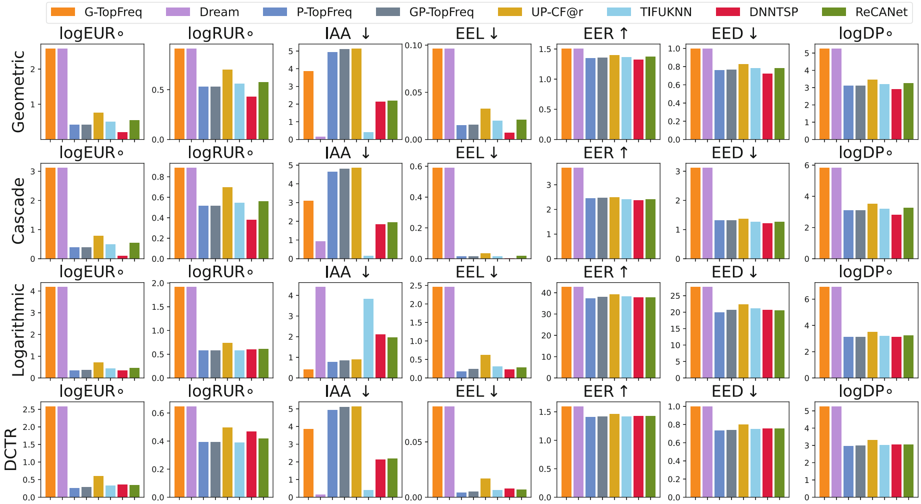


Fig. 2. Item fairness with varying position weighting models. Each column represents an item fairness metric. Each row represents a position weighting model.

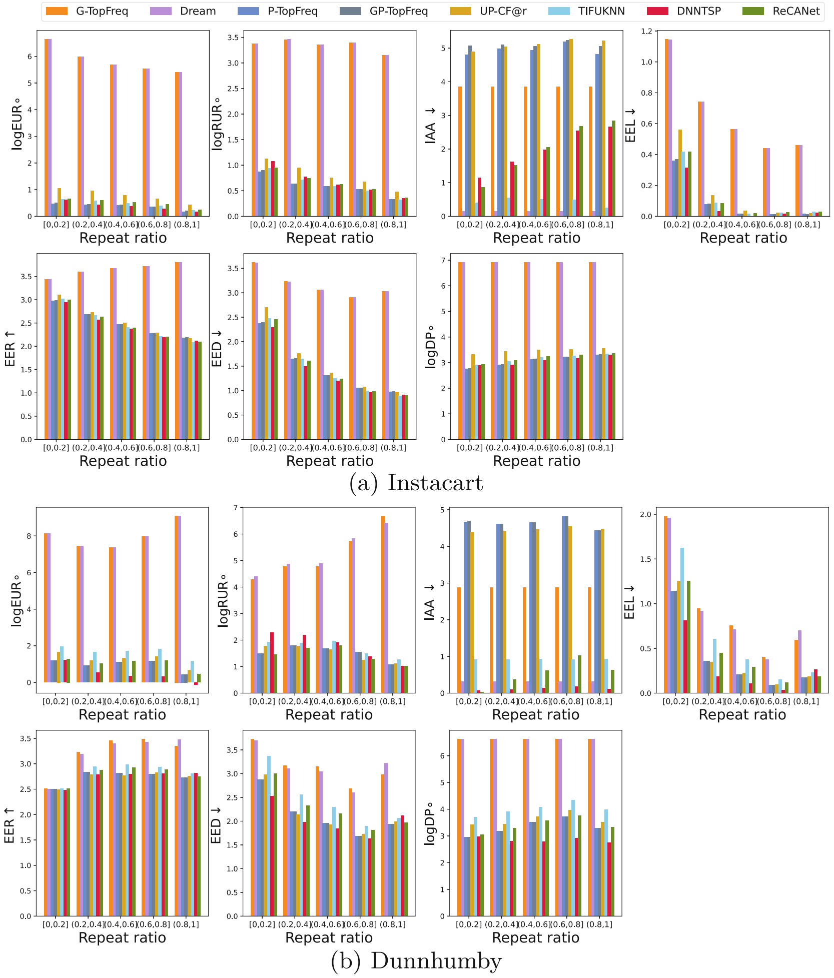
non-monotonic changes; and these metrics reorder the NBR methods. (iii) In line with [26], we see that EEL, EED, and EER are stable with varying basket sizes.

Position Weighting Models. To answer **RQ3**, we perform experiments with different position weighting models (in Table 4) for each fairness metric and report the results of Instacart[5] in Fig. 2. The parameters are assigned to the default values: patience parameter $\gamma = 0.5$, and a stopping probability of 0.5.[6] We summarize our observations below: (i) Different from [26], we observe high sensitivity of IAA on the three datasets for different position weighting models. (ii) Except for IAA, other metrics maintain the order of NBR methods across different position weighting models. Also, some slight ordering adjustments happen among methods with similar fairness values. For example, for logEUR in Fig. 2, the relative fairness order of TIFUKNN and DNNTSP is changed when the position weight model is changed from Logarithmic to DCTR. (iii) We find that the exposure values for each position weighting model is quite different. Since these metrics capture the gap between exposure and relevance, this could explain why their values change significantly in Fig. 2 when using different position weighting models.

Repeat Purchase Behavior. To answer **RQ4**, we group users into five subgroups based on their repeat consumption behavior. Here, *repeat ratio* is defined as the proportion of repeat items in the ground-truth basket for each user [19]. Hence, we create five

[5] We observe a similar trend on the Dunnhumby and TaFeng datasets. Because of space limitations, we only report the results on the Instacart dataset.

[6] The Geometric and Rank-biased precision (RBP) share the same formula under this parameter setting. Therefore, we only report the results obtained by the Geometric weighting model for fairness metrics.



(a) Instacart

(b) Dunnhumby

Fig. 3. Item fairness under different repeat ratios.

subgroups of users with a repeat ratio of $[0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, 1.0]$, respectively. We calculate fairness metrics for each subgroup to test the sensitivity of fairness metrics to different degrees of repeat purchase behavior and report the results in Fig. 3. We can observe that: (i) The fairness metrics show clear differences for different user groups, indicating that repeat purchase behavior does affect item fairness. (ii) The pattern on the Instacart dataset is relatively clear (Fig. 3a). logEUR, logRUR, and EEL consider user groups with higher repeat ratios fairer because they depend highly on the utility of the NBR methods. As NBR utility increases for higher repeat ratios [19], fairness for higher repeat ratios increases too. (iii) EER also belongs to the notion of equal opportunity, but these repeat-biased methods become more unfair as the repeat ratio increases, which is inconsistent with logEUR, logRUR, and EEL.

(iv) IAA is only highly related to the predicted relevance, which is not affected by the utility of NBR methods. IAA values for explore-biased methods (GTopFreq and Dream) are stable. Only DNNTSP and ReCANet become more unfair when the repeat ratio increases, while IAA values for other methods fluctuate. (v) EED and logDP only measure whether the distribution of the popular and unpopular groups is uniform. From logDP, we see that, except for GTopFreq and Dream, the popular group gains more exposure when the repeat ratio increases. Since GTopFreq and Dream only recommend popular items to users, their logDP measurements do not change. However, EED indicates fairer for all methods as the repeat ratio goes up, which is contradictory to logDP. The position weighting model of EED is cascade, giving a discount to the exposure of correctly predicted items. User groups with higher repeat ratios have more correctly predicted items; therefore, the EED values decrease even though there are actually more spots given to popular items for user groups with higher repeat ratios. (vi) Patterns on the Dunnhumby (Fig. 3b) and TaFeng datasets[7] are more complex. Some metrics do not change monotonically as the repeat ratio goes up. For instance, EEL and EED agree that the recommendation for user group with repeat ratio $(0.6, 0.8]$ demonstrates the best item fairness.

5 Conclusion

In this paper we have reproduced the fairness metrics implementation and empirical experiments in [26] to investigate whether the lessons about fairness metrics can be generalized to next basket recommendation. Specifically, we measure the item fairness of the NBR methods and find that these metrics give different fairness rankings of the NBR methods. However, most of the metrics agree that repeat-biased methods are fairer than explore-biased methods. Different from the observations in [26], IAA is the most sensitive metric to both basket size and position weighting models. Finally, we analyze how repeat purchase behavior affects item fairness from the perspective of both equal opportunity and statistical parity. Above all, we recommend using EEL and EED for NBR since they show high robustness towards parameter configuration and various position weighting models, and can measure fairness for multiple groups.

Our work confirms that fairness metrics show different patterns of sensitivity for different information access systems due to the characteristics of the scenarios and subtle differences in metrics implementation. For the sake of rigor, we suggest testing the sensitivity of fairness metrics for every specific scenario, following the evaluation framework used by us and by Raj and Ekstrand [26] to ensure the employed metrics are reliable.

This paper is the first attempt to study item exposure fairness in NBR domain, which provides a reference for the optimization direction of subsequent NBR methods. We have mainly considered group fairness grouped by popularity, but it is worth examining other ways of grouping, such as by brand or category. From a practical perspective, equal opportunity is a safe and reasonable optimization goal, however, we cannot conclude whether statistical parity is applicable to NBR since it is not reasonable to assign

[7] The pattern on the TaFeng dataset is similar to that on the Dunnhumby dataset. Because of space limitations, we report the results on the TaFeng dataset in the repository.

equal exposure to popular and unpopular groups in practical grocery shopping scenarios. Nevertheless, future work should investigate the ideal exposure distribution of the two groups.

Acknowledgments. We thank our reviewers for their valuable feedback.

This research was supported by the China Scholarship Council under grant nrs. 202206290080 and 20190607154, Ahold Delhaize, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture, and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

1. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The unfairness of popularity bias in recommendation. In: 13th ACM Conference on Recommender Systems, RecSys 2019 (2019)
2. Ariannezhad, M., Jullien, S., Li, M., Fang, M., Schelter, S., de Rijke, M.: ReCANet: a repeat consumption-aware neural network for next basket recommendation in grocery shopping. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1240–1250 (2022)
3. Ariannezhad, M., Li, M., Jullien, S., de Rijke, M.: Complex item set recommendation. In: SIGIR 2023: 46th international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3444–3447, ACM (July 2023)
4. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: amortizing individual fairness in rankings. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 405–414 (2018)
5. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: Proceedings of the 2008 international conference on web search and data mining, pp. 87–94 (2008)
6. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 275–284 (2020)
7. Ekstrand, M.D., Carterette, B., Diaz, F.: Distributionally-informed recommender system evaluation. ACM Transactions on Recommender Systems (2023)
8. Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness in information access systems. Found. Trends Inf. Retr. **16**(1–2), 1–177 (2022)
9. Faggioli, G., Polato, M., Aiolli, F.: Recency aware collaborative filtering for next basket recommendation. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 80–87 (2020)
10. Ge, Y., et al.: Towards long-term fairness in recommendation. In: Proceedings of the 14th ACM international conference on web search and data mining, pp. 445–453 (2021)
11. Goyani, M., Chaurasiya, N.: A review of movie recommendation system: limitations, survey and challenges. ELCVIA: Electron. Lett. Comput. Vision Image Anal. **19**(3), 0018–37 (2020)

12. Hu, H., He, X.: Sets2sets: learning from sequential sets with neural networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1491–1499 (2019)
13. Hu, H., He, X., Gao, J., Zhang, Z.L.: Modeling personalized item frequency information for next-basket recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1071–1080 (2020)
14. Katz, O., Barkan, O., Koenigstein, N., Zabari, N.: Learning to ride a buy-cycle: a hyper-convolutional model for next basket repurchase recommendation. In: Proceedings of the 16th ACM Conference on Recommender Systems, pp. 316–326 (2022)
15. Kowald, D., Schedl, M., Lex, E.: The unfairness of popularity bias in music recommendation: a reproducibility study. In: Jose, J.M., et al. (eds.) Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II, pp. 35–42. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_5
16. Le, D.T., Lauw, H.W., Fang, Y.: Correlation-sensitive next-basket recommendation. In: the Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 10–18 (2019)
17. Li, M., Ariannezhad, M., Yates, A., de Rijke, M.: Masked and swapped sequence modeling for next novel basket recommendation in grocery shopping. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 35–46 (2023)
18. Li, M., Ariannezhad, M., Yates, A., de Rijke, M.: Who will purchase this item next? Reverse next period recommendation in grocery shopping. ACM Trans. Recomm. Syst.**1**(2), Article 10 (June 2023)
19. Li, M., Jullien, S., Ariannezhad, M., de Rijke, M.: A next basket recommendation reality check. ACM Trans. Inform. Syst. **41**(4), 1–29 (2023)
20. Li, X., et al.: Mitigating frequency bias in next-basket recommendation via deconfounders. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 616–625, IEEE (2022)
21. Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: Proceedings of the Web Conference 2021, pp. 624–632 (2021)
22. Li, Y., et al.: Fairness in recommendation: a survey. ACM Transactions on Intelligent Systems and Technology (2022)
23. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inform. Syst. (TOIS) **27**(1), 1–27 (2008)
24. Naumov, S., Ananyeva, M., Lashinin, O., Kolesnikov, S., Ignatov, D.I.: Time-dependent next-basket recommendations. In: European Conference on Information Retrieval, pp. 502–511, Springer (2023)
25. Qin, Y., Wang, P., Li, C.: The world is binary: contrastive learning for denoising next basket recommendation. In: Proceedings of the 44th International ACM Sigir Conference on Research and Development in Information Retrieval, pp. 859–868 (2021)
26. Raj, A., Ekstrand, M.D.: Measuring fairness in ranked results: an analytical and empirical comparison. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 726–736 (2022)
27. Sapiezynski, P., Zeng, W., E Robertson, R., Mislove, A., Wilson, C.: Quantifying the impact of user attention fair group representation in ranked lists. In: Companion proceedings of the 2019 World Wide Web Conference, pp. 553–562 (2019)
28. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2219–2228 (2018)
29. Sun, W., Xie, R., Zhang, J., Zhao, W.X., Lin, L., Wen, J.R.: Generative next-basket recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 737–743 (2023)

30. Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: A survey on the fairness of recommender systems. ACM Trans. Inform. Syst. **41**(3), 1–43 (2023)
31. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–6 (2017)
32. Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T.: A dynamic recurrent model for next basket recommendation. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 729–732 (2016)
33. Yu, L., Sun, L., Du, B., Liu, C., Xiong, H., Lv, W.: Predicting temporal sets with deep neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1083–1091 (2020)
34. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA*IR: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1569–1578 (2017)