Towards Measuring Fairness in Grid Layout in Recommender Systems

AMIFA RAJ, People and Information Research Team

Boise State University, USA

MICHAEL D. EKSTRAND, People and Information Research Team

Boise State University, USA

There has been significant research in the last five years on ensuring the providers of items in a recommender system are treated fairly, particularly in terms of the exposure the system provides to their work through its results. However, the metrics developed to date have all been designed and tested for linear ranked lists. It is unknown whether and how existing fair ranking metrics for linear layouts can be applied to grid-based displays. Moreover, depending on the device (phone, tab, or laptop) users use to interact with systems, column size is adjusted using column reduction approaches in a grid-view. The visibility or exposure of recommended items in grid layouts varies based on column sizes and column reduction approaches as well. In this paper, we extend existing fair ranking concepts and metrics to study provider-side group fairness in grid layouts, present an analysis of the behavior of these grid adaptations of fair ranking metrics, and study how their behavior changes across different grid ranking layout designs and geometries. We examine how fairness scores change with different ranking layouts to yield insights into (1) the consistency of fair ranking measurements across layouts; (2) whether rankings optimized for fairness in a linear ranking remain fair when the results are displayed in a grid; and (3) the impact of column reduction approaches to support different device geometries on fairness measurement. This work highlights the need to use layout-specific user attention models when measuring fairness of rankings, and provide practitioners with a first set of insights on what to expect when translating existing fair ranking metrics to the grid layouts in wide use today.

CCS Concepts: • **Information systems** → **Evaluation of retrieval results**.

1 INTRODUCTION

Recommender systems may induce unfair distribution of exposure across items and their providers on either individual or group basis, often reflecting societal or historical bias such as prioritizing items from certain races or genders. An "equality of opportunity" goal [23] ensures that two providers whose items are equally useful to a user's information need have the same opportunity to be exposed to users, but systems do not always meet this criteria, instead providing *disparate exposure* [10]. There are several metrics to measure fairness of exposure (or related constructs) in ranked lists [23], but they are designed for linear — usually vertical — layouts. However, many systems use other ranking layouts such visual grids or voice responses. Grid layouts (figure 1(c)) are particularly popular for streaming media platforms and image search, but also appear elsewhere; there has been little work to determine how to measure group fairness in such layouts, or how to measure fairness when the system may use different layouts in different contexts or device. It can be problematic to measure grid-layout fairness by simply mapping the grid positions to a linear layout and using existing metrics, because user attention to items in different positions varies between layouts [7]. For the same set of recommended items, user attention varies depending on how the items are being displayed, affecting item exposure and therefore the fairness of that exposure. Using fair ranking metrics without taking layout-specific user browsing behaviour into consideration may provide unreliable and erroneous results.

Further, based on the device (phone, tablet, TV, laptop, etc.) used to interact with a system, the geometry of grid layouts varies, often re-ranking the list as the number of available columns changes. There are also multiple methods for adjusting the layout: for example, when moving from a wider to a narrower screen, some systems *truncate* the list at the right-side while others *re-wrap* the entire list. The impact of these layout adjustments on fairness scores is unknown. In

summary, researchers and developers using grid layouts have little to work with when trying to reason about how the system layouts affect equity of exposure or how to apply the various metrics that have been developed to this setting.

In this paper, we seek to fill this gap and broaden the applicability of fair ranking metric research by extending fair ranking metrics to grid layouts, providing the first (to our knowledge) study of metrics for this widely-used but under-studied paradigm. We observe what happens to group fairness for a list of recommended items with the change of layouts by answering the following research questions:

RQ1. Do fairness measurements remain consistent across layouts?

RQ2. Do rankings optimized for fairness in linear layouts remain fair in grids?

RQ3. How do provider-side group fairness scores change as grid size changes?

RQ3.a. Does the fair ranking metric score change when the grid layout is truncated or re-wrapped?

RQ3.b Does the change in fairness score with column-size reduction remain consistent across reduction approaches?

The main contributions of this work are to:

- Describe various types of layouts that are often used to display recommended items.
- Incorporate grid browsing models into fair ranking metrics to derive fair grid metrics
- Provide insights on fairness score consistency and applicability across layouts.
- Describe the impact of column reduction approaches on fairness scores within a grid layout.

(a) Linear Vertical Layout (b) Linear Horizontal Layout (c) Wrapped Grid Layout (d) Multi-list Grid Layout

Fig. 1. Various Types of Layouts

2 BACKGROUND AND RELATED WORK

This work draws from a line of work on fair RS that we review here and browsing models in sec 3.4.

Recommender systems often present recommended results in top-N ranked order based on relevance to user information preference. Thus systems *expose* recommended items along with their providers through ranked lists and these ranked lists can be represented in *linear* (figure 1(a)) or *grid* (figure 1(c)) layouts.

It is not possible for items with the same relevance to get the same position in a single ranked lists, and a small change of relevance causes item position to vary [27], thus affecting user attention they receive. RS can cause *disparate exposure* based on provider group association while distributing exposure across relevant items. User attention is not uniformly distributed across items in a ranked list and users tend to interact more with items at the top positions, which causes both economic and reputational disadvantage to the items (and their providers) at lower-ranked positions. Thus, items with similar merit will not receive similar benefits due the position difference or *disparate exposure*.

Table 1. Summary of notation.

$d \in D$	document or item
$q \in Q$	request (user or context)
L	ranked results of N documents from D
$L^{-1}(i)$	the document in position i of linear (1-column) layout
$L(d)$	rank of document d in linear layout
$\text{row}(d)$	row number of document d in grid layout
$L^{-1}(\text{k}, \cdot)$	items in kth row in grid layout
$L^{-1}(\text{k}, \text{c})$	items in row k and column c in grid layout
$y(d\|q)$	relevance of d to q
g	number of groups
$\mathcal{G}(d)$	group alignment vector
$\mathcal{G}(L)$	group alignment matrix for documents in L
$\mathcal{G}^{+}(L)$	set of documents in protected group in L
$\mathcal{G}^{-}(L)$	set of documents non-protected group in L
$\hat{\mathbf{p}}$	target group distribution
$\mathbf{a}_L$	attention vector for documents in L
$\mathbf{a}_L(d)$	position weight of d in L
$\boldsymbol{\epsilon}_L$	the exposure of groups in L

The focus of this work is on **provider-side group fairness** in RS ranking ensuring that different groups of item providers do not experience unjustified discrepancies in the exposure of their content on the basis of their sensitive attributes, such as gender or ethnicity. Several metrics have been proposed to measure provider-side group fairness in ranking. The broader goal of these metrics is to measure system's ability to allocate fair exposure across item providers based on their group membership and thus they measure exposure discrepancy across groups in ranking. Yang and Stoyanovich [33], Zehlike et al. [34], and Sapiezynski et al. [25], among others proposed metrics that measure group fairness for providers in a single ranking. Without considering relevance information these metrics measure fairness as *statistical parity* where item position should not be affected by group membership. Among these metrics Sapiezynski et al. [25] use *position weights* derived from a user attention model to measure the fairness of item exposure. Biega et al. [4], Singh and Joachims [27], and Diaz et al. [10] proposed metrics that considered relevance information in fairness measurement and measure fairness as *equal opportunity* where exposure or attention should be proportional to relevance. These metrics measure fairness in sequences or distributions over rankings since it is not possible to achieve fair exposure in a single ranking. All these metrics also consider position weight in their fairness measurement. Beutel et al. [3] and Narasimhan et al. [21] took a different approach and measure fairness by considering pairwise ordering.

Kuhlman et al. [18] provided a comparative analysis of selected metrics that can measure statistical parity and Raj and Ekstrand [23] provided a comprehensive and comparative analysis of existing metrics that are suitable to measure provider-side group fairness in ranking showing the conceptual similarities and differences among the metrics. Raj and Ekstrand [23] identified metrics that considers user attention changing behaviour with item positions to determine position weight in ranking. Various user browsing models are considered to demonstrate user browsing behaviors in ranking and their study showed that metrics show sensitivity towards the choice of user browsing models.

3 PROBLEM FORMULATION

In this work, we consider a recommender system that recommends n items $d_1, d_2, \ldots, d_n \in D$ in response to information requests from users $q_1, q_2, \ldots, q_m \in Q$ based on their relevance to the request $y(d|q)$ and presents the results in a layout L (either 1-column, as in a classical linear layout, or a multi-column layout). Documents are associated with producers or providers who in turn can be associated with demographic attributes identifying them with one or more of g groups. We model group membership of documents with group alignment vector $\mathcal{G}(d) \in [0, 1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$) forming a

distribution over groups; this allows for mixed, partial, or uncertain membership in an arbitrary number of groups. Table 1 summarizes the notation used in this paper.

3.1 Ranking Layouts

Without loss of generality, we treat recommendation and layout as a multi-stage process: the system first scores and ranks items for the user (either deterministically or with a stochastic policy [10]), and then displays that ranking in a layout. In this work, we consider layouts in $r \times c$ grids, where r is the number of rows and c the number of columns; this encapsulates at least four distinct models. One such family of layouts comprise the familiar **linear layouts** where items are displayed in a single linear list. These come in two varieties

Vertical Ranking Model Items are displayed in a multi-row single-column ranked list ($r \times 1$, see Figure 1(a)). Users generally see items from top to bottom. The layout may be split into multiple pages.

Horizontal Ranking Model Items are displayed in single-rows with multiple column lists following $1 \times c$ pattern, see figure 1(b). Users see items from left to right.

In **grid layout**, items are displayed in multiple rows and columns ($r \times c$). These also come in multiple varieties:

Wrapped Grid Items are displayed as a single ranking in an $r \times c$ grid, without being categorized into groups, see figure 1(c). The grid is formed by displaying the items in order horizontally and starting a new row when the display runs out of space.

Multi-ranking Grid Items are displayed in multiple rows, often based on categories or recommendation sources, and each row consists of a ranked list of items. In figure 1(d), recommended items are categorized by genre which may facilitate users to find them from their preferred categories.

We focus on **wrapped grid** layout in this work due to the better availability of browsing and attention models for this problem setting. Further work is needed to provide usable models of user browsing behavior with multi-ranking grids before we can attempt to measure their fairness.

3.2 Fair Ranking Metrics

We follow the recommendations of Raj and Ekstrand [23] and study two metrics: *Attention-Weighted Rank Fairness* [AWRF$_\Delta$, 25] to measure *statistical parity* in single ranking (averaging over multiple rankings to measure overall system fairness), and *Expected Exposure Loss* [EEL, 10] to measure equal opportunity in sequences of rankings. These metrics measure the distribution of exposure (based on estimated user attention) across provider groups to measure the fairness of rankings. They represent user attention with a *position weight* assigned to each document in a ranking.

Both metrics rely on a model of user attention (estimating the attention a user is likely to give to items at different positions in ranking) in order to measure fairness; it is important to know how users browse and interact with different positions in the ranked layout. Several studies have used user eye gaze tracker to study user browsing behavior [11, 26, 31, 35]. Some studies used user click behavior to infer browsing behavior of users [12, 31] with respect to ranking positions. Simple models of user browsing behavior, commonly used in information retrieval metrics and described in the next sections, determine these weights based on items' position in the ranking (along with other information, such as the relevance of preceding documents).

AWRF$_\Delta$ is suitable to measure provider-side fairness in single ranking and it measures the difference between group exposure and configurable *population estimator* (representing the ideal distribution of exposure over groups) using a distance function Δ. The exposure for each groups ϵ_L is derived from the attention vector and the group

alignment matrix ($\boldsymbol{\epsilon}_{\mathrm{L}} = \mathcal{G}(\mathrm{L})^{\mathrm{T}}\mathbf{a}_{\mathrm{L}}$) which aggregates the attention given to items of each group in proportion to their group membership as represented by the alignment vector:

$$\mathrm{AWRF}_{\Delta}(L) = \Delta(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}}) \tag{1}$$

EEL is suitable for *stochastic* ranking policy where fairness is measured over user-dependant distribution over rankings $\rho(L|q)$ since it is not possible to achieve equal exposure in single ranking [10]. It can be drawn as distribution over rankings $L_1, L_2, \ldots, L_{\tilde{n}}$ from the distribution over requests $\rho(q)\pi(L|q)$ [23]. EEL uses available relevance information to derive a *target exposure* $\boldsymbol{\epsilon}_\tau$, based on an ideal policy τ where relevant items are sorted in non-decreasing order in ranking and exposure is fairly distributed across the relevant items. Using the $\boldsymbol{\epsilon}_L$ of each ranking, the system exposure is derived as $\boldsymbol{\epsilon}_\pi = \sum_L \pi(L|q)\boldsymbol{\epsilon}_L$. EEL is computed as the squared Euclidean distance between system exposure $\boldsymbol{\epsilon}_\pi$ and target exposure $\boldsymbol{\epsilon}_\tau$:

$$\mathrm{EEL} = \|\boldsymbol{\epsilon}_\pi - \boldsymbol{\epsilon}_\tau\|_2^2 \tag{2}$$

Since these metrics compute fairness as a distance from the target distribution regardless of layout, it is meaningful to directly compare fairness scores between layouts for the same test data.

3.3 Linear Browsing Models

In linear ranking, users typically browse the list from top to bottom [8], with the probability that they will continue (and thus view more items) decreasing as they move down the list. There are a variety of models of this scanning behavior with decaying attention; *cascade* and *geometric* are commonly-used click models to estimate user interaction probability with ranking positions. These models have been employed to construct evaluation metrics to measure utility [1, 5, 6, 20] or item exposure [4, 10, 25] in rankings.

Moffat and Zobel [20] proposed the *rank-biased precision* (RBP) evaluation metric to weight precision based on user attention to different ranking positions. This metric used a geometric browsing model with a *continuation probability* α to estimate the probability of users moving to the next item (position) or stopping (click) at that position; the visiting probability exponentially decreases with ranking positions. Biega et al. [4] proposed a modified version where the position weight decays geometrically with each position having the same probability of being stopped (clicked). In this model, the visiting probability of item d in position $L(d)$ is determined by:

$$P_{\text{geometric}}[V_d] = \alpha^{L(d)} \tag{3}$$

Craswell et al. [8] proposed the *cascade* click model where users will view position i if they have skipped items before that position, and whether users will click or skip a position depends on the relevance of the item in that position and the relevance of items in previous positions. Chapelle et al. [6] proposed a cascade-based metric *expected reciprocal rank* (ERR) by extending the cascade model to include the probability of users terminating the entire process as an *abandonment* probability that decays geometrically. In the cascade model, users will visit item d if they did not stop at any position before that item in the ranked list which is determined by item relevance. The continuation probability α is now a function of relevance, and the probability of visiting d is given by:

$$P_{\text{cascade}}[V_d] = \prod_{j \in [0, L(d))} \alpha\left(y\left(L^{-1}(j)\middle|q\right)\right) \tag{4}$$

Table 2. Parameters of Weighting Models for computing $a_L(d)$ and the range of parameter values

Parameters	Values	Browsing Models	Default Values
Skipping Probability γ	{0.1, 0.2, ..., 0.9}	Row Skipping	0.5
Continuation Probability α	{0.1, 0.2, ..., 0.9}	Cascade Geometric	0.5
Slow parameter β	{1.1, 1.2, ..., 2.0}	Slower Decay	1.9

3.4 Grid-based Browsing Models

Users do not interact with grid displays the same way they interact with linear displays and several studies have been performed to understand how users allocate attention to different items in grid layouts.

3.4.1 Existing Literature on User Browsing Behavior in Grid Layouts. Tatler [29] observed that users show tendency of *central fixation* where they tend to put more attention on the middle of the screen than on the edges, Djamasbi et al. [11] found that users usually focus on results located at the top left-hand side and proceed in an *F-shaped* reading pattern, but the viewing pattern varies depending on task, content, and complexity of web pages. The eye-tracking study of Zhao et al. [35] also observed an *F-pattern* in user interaction with grid-based recommendations but the pattern can vary depending on task Shrestha and Lenz [26] emphasized on the need of considering page content while understanding user viewing patterns. Xie et al. [31, 32] performed eye-tracking studies in grid-based image search results and observed the *middle bias* pattern. Moreover, in grid-view, user attention decreases at a slower rate than in linear layouts (*slower decay*) and users often jumps to results after skipping rows (*row skipping*).

The studies mentioned above mostly focus on understanding user viewing patterns in grid-based interfaces with the goal of providing and measuring user satisfaction. There is limited research work concerning fairness issues when results are displayed in grid layout. Guo et al. [16] proposed de-biasing techniques for grid-based product search result pages in e-commerce systems; consistent with the studies above, they observed that user attention follows *row skipping* and *slower decay* while interacting results in grid layout. Balyan et al. [2] emphasized on item meta information in user viewing behavior in grid-based e-commerce search results.

3.4.2 Adapting Browsing Models to Grid-Based User Behavior. Since the previous studies showed that user attention varies between applications depending on task, domain, device, and details of the layout, considering multiple viable models from existing literature will provide insights useful to researchers and practitioners in various contexts, as they can apply an appropriate model for their systems. For this preliminary analysis, we will consider *row-skipping* (RS) and *slower-decay* (SD) in the context of wrapped grid layout; we leave central fixation, multi-list rankings, and incorporating multiple browsing model adjustments simultaneously to future work.

We adapt both the *cascade* and *geometric* browsing models to account for *row-skipping* (RS) and *slower-decay* (SD). Table 2 shows the parameters and range of values we consider to measure attention weight of items in ranking. For *row skipping* behavior, the visiting probability of item d at row(d) and ranking position $L(d)$ depends on the skipping probability of a row γ; for each of the k rows before row(d), the user either continued through that row, or skipped it with probability γ. If user visited items in a row, that implies that a particular row was not skipped. With that assumption, visiting probability of item d in cascade-based row-skipping model considering relevance:

$$P_{RS(\text{cascade})}[V_d] = \left[\prod_{k=0}^{\text{row}(d)} (1-\gamma) \prod_{i \in L^{-1}(\text{k},\cdot)} \alpha(y(L^{-1}(i)|q)) + \prod_{k=0}^{\text{row}(d)} \gamma \right] \prod_{i \in \text{row}(d)} \alpha(y(L^{-1}(i)|q)) \tag{5}$$

The visiting probability of item d in geometric-based row-skipping model is given by[1]:

$$P_{RS(\text{geometric})}[V_d] = \left[\prod_{k=0}^{\text{row}(d)} (1-\gamma) \prod_{i \in L^{-1}(\text{k},\cdot)} \alpha + \prod_{k=0}^{\text{row}(d)} \gamma \right] \prod_{i \in \text{row}(d)} \alpha \tag{6}$$

With the *slower-decay* browsing behavior, visiting probability of items across a row in a grid layout decays more slowly than in a vertical linear list, but jumps when the user moves to the next row. This is modeled by a decay parameter β to modify the continuation probability for items in ranked results based on the row in which they appear. The visiting probability of item d in cascade-based slow-decay model is:

$$P_{SD(\text{cascade})}[V_d] = \min(\beta^{\text{row}(d)} \prod_{i=[0,L(d)]} \alpha(y(L^{-1}(i)|q)), 1) \tag{7}$$

The geometric visiting probability of item d with slower decay is (derived by [16]):

$$P_{SD(\text{geometric})}[V_d] = \min(\beta^{\text{row}(d)} \prod_{i=[0,L(d)]} \alpha, 1) \tag{8}$$

3.5 Changing Grid Layouts

Based on the device the users use to interact with the system, grid layout can be converted into a size suitable for a particular device using two different approaches: *truncation*, where each row is truncated and item off-screen are no longer displayed, and *re-wrapping*, where the rows are re-wrapped so the items that would be off-screen are moved to the next row. These approaches may differ in their influence on the fairness of the resulting display. To observe the impact of column size and column reduction approaches on group fairness score, we change the column size for a given grid ranking using both *truncation* and *re-wrap* approaches.

4 EXPERIMENTAL SETUP

Our central goal is to understand how measurements and optimizations for classical linear rankings apply to grid layouts, both to apply existing methods and to identify where further research is needed to support fairness in these widely-used layouts. To answer our research questions, we conduct several experiments by implementing the metrics with adaptations for user behavior in grid layouts and using them to measure outputs in real-world datasets.

4.1 Dataset

We use two user-book interaction datasets from *GoodReads* [30] and *Amazon* [19], integrated with the PIReT Book Data Tools[2] [14] to obtain author metadata. Table 3 shows the summary of the datasets. For both datasets, we generate 1000 personalized book recommendations for 5000 users using four collaborative filtering algorithms: user-based (UU [17]), item-based (II [9]), matrix factorization (WRLS [28]), and Bayesian Personalized Ranking (BPR [24]), as configured by Ekstrand and Kluver [14]. We used *Lenskit for Python* [13] to generate recommendations using binary implicit feedback for items. Author gender identity is the sensitive attributes for our experiments. Due to limitations of the underlying data set [15], this is a discrete but possibly unknown binary gender attribute; we acknowledge the importance of more faithful representations of gender in research [22], and the metrics that we study can all be used with a larger set of gender identities as well as mixed or partial membership when such data can be obtained. Both datasets have incomplete

[1]This model is derived by [16]) where they referred the model as *cascade click model*. However, in our paper, we referred the model as *geometric* to keep the conceptual consistency.
[2]https://bookdata.piret.info

Table 3. Summary of experiment data with nDCG score

Dataset	Data Statitstics			Group Sizes		nDCG			
	#Users	#Items	#Test Users	$\lvert\mathcal{G}^+\rvert$	$\lvert\mathcal{G}^-\rvert$	II	UU	WRLS	BPR
Amazon	8,026,324	2,268,142	5000	217032	490953	0.08	0.13	0.10	0.03
GoodReads	870,011	1,096,636	5000	177359	282857	0.23	0.24	0.26	0.13

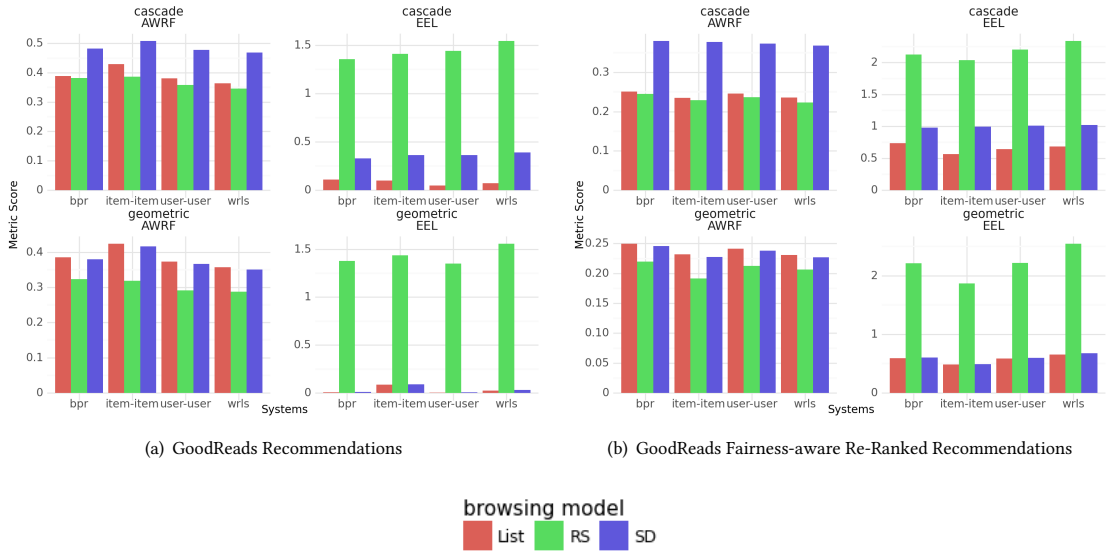
relevance judgements and incomplete group labels. We follow common practice and consider documents without relevance data as non-relevant, and treat missing group labels as a separate unknown category in our experiments.

4.2 Methodology

Across several different scenarios, we measure changes in the fairness scores themselves, as well as changes in the ranking of systems (which system is measured to be most fair), to understand the impact of layouts and browsing models on fairness measurement.

RQ1. To observe the consistency of fair ranking measurements across layouts, we represent the recommended items in linear-vertical layout and 5-column wrapped grid layout and measure fairness using the metrics in their default parameter settings. We implement AWRF_Δ and *EEL* with the modified user attention models to account for wrapped grid layout. The metric score comparison shows how the fairness scores change with the choice of layouts.

RQ2. To better understand the fairness score differences across layouts, we identify if fairness remains consistent across layouts — whether a ranked result optimized to be fair for a certain layout remains fair for other layouts. We apply the GreedyEQ group-fair reranking technique from [15] to recommendation results to generate fairness-optimized ranked lists; we then render these rankings into 5-column grid layouts and measure group fairness in both linear and grid layouts. This experiment shows the persistence of fairness scores of a ranked list across layouts.

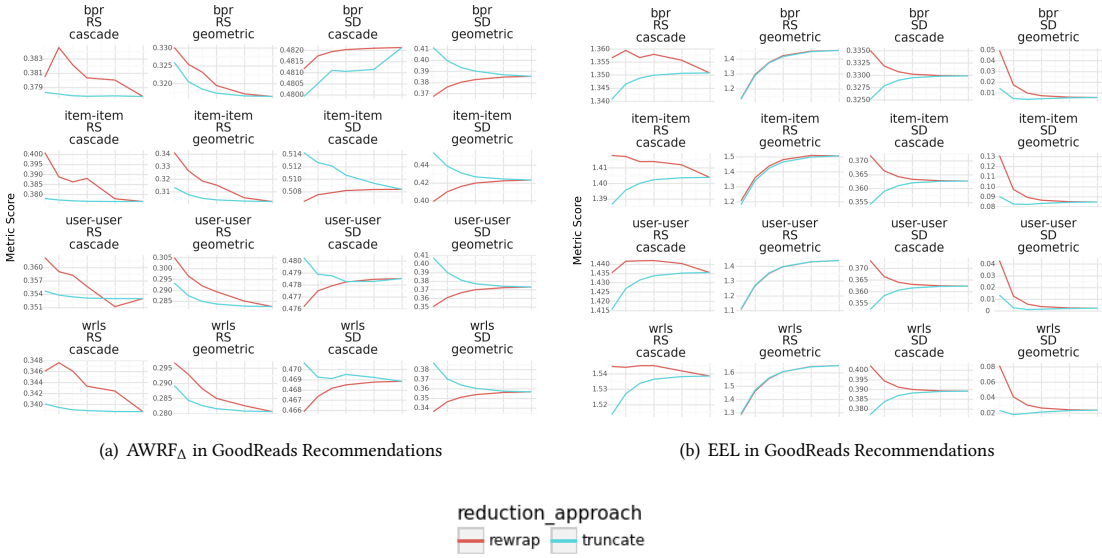


(a) GoodReads Recommendations

(b) GoodReads Fairness-aware Re-Ranked Recommendations

Fig. 2. Metrics results with the change of weighting strategy

RQ3. As noted in Section 3.5, column reduction can be done by either *truncating* or *re-wrapping* the rows to fit the user's current screen which may have different impacts on the fairness scores of system outputs. Further, fairness scores may change as column size changes regardless of approach. We represent the set of recommended items in grid layout changing the column size in 10, 8, 6, 5, 4, 3 using both truncation and re-wrap approaches. To see the impact of column size on group fairness score and the fairness score consistency across column-reduction approaches, we compare fairness scores across column sizes and across the reduction approaches.



(a) $AWRF_{\Delta}$ in GoodReads Recommendations

(b) EEL in GoodReads Recommendations

Fig. 3. Metrics results with the change of column sizes across column reduction approaches

5 RESULTS AND DISCUSSION

We now present the results of our investigation into the behavior of fair ranking measurements applied to grid layouts. We observe similar results for both datasets and due to space limitations we show results from GoodReads dataset.

5.1 RQ1: Do fairness measurements remain consistent across layouts?

Figure 2(a) shows the fair ranking metric scores change with layouts and within grid layout with the change of browsing behaviors (*row-skipping*, *slower-decay*). $AWRF_{\Delta}$ score varies across grid adjustments to browsing models, keeping the same order of systems, but the *cascade* and *geometric* browsing models rank systems in a different order. *EEL* scores with *row-skipping* model notable vary; this shift is significantly greater than the shift seen in $AWRF_{\Delta}$.

Implications. From *RQ1*, we have following observations:

- Fair raking metric scores are highly dependent on layout and user browsing model.
- Within a layout, metric score further varies across user browsing behavior.

- Since user attention is one of the required components of $AWRF_\Delta$ and EEL implementation and user attention for ranking positions is determined by user browsing behavior, it is important to consider accurate browsing model while applying these metrics.

5.2 RQ2: Do rankings optimized for fairness in linear layouts remain fair in grids?

From Figure 2(b), we see that $AWRF_\Delta$ scores are consistent across layouts specifically with *geometric* browsing model. *EEL* score for a fairness optimized ranking can vary across layouts depending on user browsing models. Within grid layout, *EEL* with the *row-skipping* browsing model provides different fairness scores and rankings than *slower-decay*.

Implications. From *RQ2*, we made following observations:

- A ranking that is fair in linear layout can be represented as unfair depending on the assumed user browsing behavior. This reinforces the need to incorporate accurate user browsing models in fairness measurement.
- Without considering layout-suitable browsing models, metrics will provide unreliable fairness scores.

5.3 RQ3: How do fairness scores change as grid size changes?

Figure 3 shows how metric score changes with column sizes and the changing pattern with column reduction approaches.

RQ3.a. Does the fair ranking metric score change when the grid layout is truncated or re-wrapped? When columns are reduced using the *truncate* approach, metrics show some stability towards column size for most of the systems. However, column size has more impact on $AWRF_\Delta$ scores than EEL with the *truncate* approach. When columns are reduced using the *re-wrap* approach, $AWRF_\Delta$ shows high sensitivity towards column sizes.

RQ3.b Does the change in group-fairness score with column size reduction remain consistent across truncation and re-wrap approach? Metric scores vary with the change of column sizes and the direction of this change is different between column reduction approaches. However, for some systems, metric scores with both column reduction approaches converges at some column sizes. In both datasets, the metrics are consistent across systems.

We do note that the *truncate* approach is primarily used with multi-list layouts in practice, while our results here are for wrapped layouts; however, finding that the use of truncation has significant effects on fairness has implications for fair layouts regardless of the initial grid layout method.

Implications. From *RQ3* we have following observations:

- Device is an important factor in measuring fairness.
- With the change of device (column size) fairness scores show high sensitivity which indicates the importance of carefully selecting column-reduction approaches while re-ranking the grid layout.

5.4 Discussion

In this work, we consider a gap in the state of the art in measuring the provider-side fairness of rankings by considering grid layouts. We apply existing fair ranking metrics in linear and grid layouts to identify their consistency across layouts. Our findings provide insights on implementation and reliability of fair ranking metrics in grid layout and provides knowledge on how metric behavior changes across ranking layouts and across column-reduction approaches within grid layouts. Our results suggest that metrics can vary in their consistency across ranking layouts ($AWRF_\Delta$ was more consistent across layouts than *EEL*). However, a metric that is consistent across layouts may not be stable across device sizes within a particular grid layout (*EEL* was more consistent across column sizes, while the consistency of $AWRF_\Delta$

metric results notably varies depending on the column-reduction approach.) Therefore, our results advise researchers and practitioners to pay close attention to ranking layout, device sizes, and column-reduction approaches while using a metric to measure fairness in ranking. Even though AWRF_Δ metric score is consistent across layouts to some extent, while using AWRF_Δ in grid layouts, practitioners should pay attention to column sizes and column-reduction approaches, whereas while using *EEL* to measure fairness in ranking, ranking layout must be taken into account but the reduction approach has less impact on the measurements.

Furthermore, our results indicate that metrics can be highly affected by user browsing behavior. Since the concept of provider-side fairness in ranking often relies on the attention in different positions, it is important to use accurate models of user attention behavior when measuring provider-side fairness in ranking. It is necessary to develop a clear and detailed understanding of user browsing behavior in order to generate valid and trustworthy fairness score using fair ranking metrics. Our work is not able to directly provide those measurements, but provides a first analysis of what to expect when applying existing measurements with the current public state of knowledge in user behavior modeling.

REFERENCES

[1] Azin Ashkan and Charles LA Clarke. 2011. On the informativeness of cascade and intent-aware effectiveness measures. In *Proceedings of the 20th International Conference on World wide web*. 407–416.

[2] Apoorva Balyan, Atul Singh, Praveen Suram, Deepak Arora, and Varun Srivastava. 2021. Using Product Meta Information for Bias Removal in E-Commerce Grid Search. *IEEE Data Eng. Bull.* 44, 2 (2021), 81–91.

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in Recommendation Ranking Through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220. https://doi.org/10.1145/3292500.3330745

[4] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 405–414. https://doi.org/10.1145/3209978.3210063

[5] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*. 903–912.

[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. 621–630.

[7] Sirui Chen, Xiao Zhang, Xu Chen, Zhiyu Li, Yuan Wang, Quan Lin, and Jun Xu. 2022. Reinforcement Re-ranking with 2D Grid-based Recommendation Panels. *arXiv preprint arXiv:2204.04954* (2022).

[8] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[9] Mukund Deshpande and George Karypis. 2004. Item-based Top-n Recommendation Algorithms. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 143–177. https://doi.org/10.1145/963770.963776

[10] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[11] Soussan Djamasbi, Marisa Siegel, and Tom Tullis. 2011. Visual hierarchy and viewing behavior: An eye tracking study. In *Human-Computer Interaction. Design and Development Approaches: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part I 14*. Springer, 331–340.

[12] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 331–338.

[13] Michael D. Ekstrand. 2020. LensKit for Python: Next-Generation Software for Recommender Systems Experiments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2999–3006. https://doi.org/10.1145/3340531.3412778

[14] Michael D. Ekstrand and Daniel Kluver. 2020. Exploring Author Gender in Book Rating and Recommendation. *User Modeling and User-Adapted Interaction* (feb 2020). https://doi.org/10.1007/s11257-020-09284-2

[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*. 242–250.

[16] Ruocheng Guo, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. 2020. Debiasing grid-based product search in e-commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2852–2860.

[17] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An Algorithmic Framework for Performing Collaborative Filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 230–237.

[18] Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*.

[19] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[20] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[21] Harikrishna Narasimhan, Andrew Cotter, Maya R Gupta, and Serena Wang. 2020. Pairwise Fairness for Ranking and Regression.. In *AAAI*. 5248–5255.

[22] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. *arXiv preprint arXiv:2301.04780* (2023).

[23] Amifa Raj and Michael D Ekstrand. 2022. 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*. 726–736.

[24] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, Quebec, Canada) *(UAI '09)*. AUAI Press, Arlington, Virginia, USA, 452–461.

[25] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595

[26] Sav Shrestha and Kelsi Lenz. 2007. Eye gaze patterns while searching vs. browsing a website. *Usability News* 9, 1 (2007), 1–9.

[27] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[28] Gábor Takács, István Pilászy, and Domonkos Tikk. 2011. Applications of the Conjugate Gradient Method for Implicit Feedback Collaborative Filtering. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) *(RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 297–300. https://doi.org/10.1145/2043932.2043987

[29] Benjamin W Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision* 7, 14 (2007), 4–4.

[30] Mengting Wan and Julian McAuley. 2018. Item Recommendation on Monotonic Behavior Chains. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 86–94. https://doi.org/10.1145/3240323.3240369

[31] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 275–284.

[32] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based evaluation metrics for web image search. In *The world wide web conference*. 2103–2114.

[33] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[34] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information Processing & Management* 59, 1 (2022), 102707.

[35] Qian Zhao, Shuo Chang, F Maxwell Harper, and Joseph A Konstan. 2016. Gaze prediction for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 131–138.