Measuring Fairness in Ranked Results

An Analytical and Empirical Comparison

Amifa Raj
People and Information Research Team
Boise State University
Boise, Idaho
amifaraj@u.boisestate.edu

Michael D. Ekstrand
People and Information Research Team
Boise State University
Boise, Idaho
michaelekstrand@boisestate.edu

ABSTRACT

Information access systems, such as search and recommender systems, often use ranked lists to present results believed to be relevant to the user's information need. Evaluating these lists for their *fairness* along with other traditional metrics provides a more complete understanding of an information access system's behavior beyond accuracy or utility constructs. To measure the (un)fairness of rankings, particularly with respect to the protected group(s) of producers or providers, several metrics have been proposed in the last several years. However, an empirical and comparative analyses of these metrics showing the applicability to specific scenario or real data, conceptual similarities, and differences is still lacking.

We aim to bridge the gap between theoretical and practical application of these metrics. In this paper we describe several fair ranking metrics from the existing literature in a common notation, enabling direct comparison of their approaches and assumptions, and empirically compare them on the same experimental setup and data sets in the context of three information access tasks. We also provide a sensitivity analysis to assess the impact of the design choices and parameter settings that go in to these metrics and point to additional work needed to improve fairness measurement.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; • **Social and professional topics → User characteristics**.

KEYWORDS

fair ranking, fairness metrics, group fairness

ACM Reference Format:
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 (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3532018

1 INTRODUCTION

Information access systems (IAS), such as search and recommender systems, often present items in response to user information needs in the form of top-N ranked lists based on relevance score and other measures of items' quality and relationships (e.g. similarity, as in maximal marginal relevance [9]).

Through these ranked lists, the system *exposes* its items (and their creators) to users, and this exposure affects what users discover, consume, and purchase. Further, this exposure is not always evenly or fairly distributed; different items or groups of items may receive *disparate exposure* when exposure is not equitably distributed to (relevant) items [11]. Disparate exposure can disadvantage content creators on either an individual or group basis. Popularity bias [1], for example, provides an advantage to creators based on their prior popularity. The system may also, however, provide greater or lesser exposure reflecting in ways that reproduce historical and ongoing social discrimination, such as by disadvantaging creators of a particular gender or race.

In the last few years, several metrics have been proposed to measure the *fairness* of rankings, some directly in search or recommendation contexts and others for more general ranking purposes such as college rankings or university admissions.

Kuhlman et al. [22] compare selected fair ranking metrics for measuring the *statistical parity* of rankings (whether they provide equal exposure to different groups), and Zehlike et al. [38] provides a thorough conceptual survey of fair ranking constructs, but there has not yet been a systematic comparison of group fairness metrics for ranked IAS outputs (where the system provides different rankings in response to different information needs — both prior comparisons focus on rankings for a single need), or direct comparisons within the same data set and experiment.

Moreover, several of the metrics have been tested primarily on small and/or synthetic data sets that are not reflective of real-world information access applications and experiments. Realistic experimental settings present challenges for applying many metrics, including incomplete data (for both relevance and group membership) and the occurrence of edge cases such as a group with no retrieved (or relevant) items. Metrics need to be robust and usable in such situations in order to be practically useful in experiments and for auditing deployed applications. Metric results may also be heavily influenced by parameter choices and experimental designs. This is an important factor to consider when choosing framework-applicable metrics because metrics which are significantly sensitive towards external factors or design choices are more complex to apply, as those decisions must be calibrated appropriately. Therefore, despite the progress in metrics for measuring fairness, both practitioners and researchers may have difficulty finding the most applicable metric for their problem setting and its requirements.

In this work, we seek to fill this gap: to provide a comparative analysis of fairness metrics in the context of information access,

SIGIR '22, July 11–15, 2022, Madrid, Spain

to better inform the community of their relative strengths and weaknesses, and facilitate both better application of existing metrics and further research to advance the state of the art in measuring ranking fairness. Our goal is not to identify universally best metrics; the essentially contested nature of fairness [29] implies such a quest is futile. Rather, we want to connect fair ranking metrics with applications by providing insight into how to measure the provider-side group fairness of the ranked outputs in actual search and recommender systems experiments using these metrics. Further, there is not a fixed ground truth to use to assess external support for a metric — there are many potential fairness objectives with varying degrees of compelling arguments. We therefore seek to inform the discussion through internal support: documenting and comparing the structure of the metrics, and their varying behaviors over real data, to assess and suggest what strengths and weaknesses we can for each metric.

Our aim is to do for provider-side group fairness what Friedler et al. [16] did for fair classification metrics; this complements the thorough conceptual survey of fair ranking constructs and interventions in a general ranking setting by Zehlike et al. [38] and Kuhlman et al. [22]. We provide a concise treatment of fair ranking metrics specifically focused on measuring fairness in information access settings, and implement these metrics in a common experimental setting to show their results on the same data, systems, and tasks. This enables us to investigate the following:

- What is needed to apply these metrics to real IAS outputs, which often have missing data (including relevance judgments and group annotations), may have highly imbalanced outputs or relevant sets, or exhibit other edge-case behavior?
- What are the actual substantive differences between these metrics, once superficial differences in framing and notation are resolved?
- What are the design decisions and parameters involved, and how sensitive are the resulting metrics to those decisions?
- What are the empirical differences in how these metrics assess the relative fairness of different recommendation algorithms or retrieval runs?

In this paper, we make four contributions:

- We describe rank fairness metrics in a unified notation for information access, identifying similarities and differences.
- We identify gaps between the conceptual form and the practicalities of applying the metrics to both search and recommender system evaluation experiments.
- We directly compare the outcomes of these metrics with the same data and experimental settings[1].
- We conduct sensitivity analysis to assess the impact of design choices and external factors on these metrics.

From our results we highlight strengths and limitations of the metrics, finding that some of them are particularly sensitive to edge cases and/or parameter settings. We conclude with recommendations for choosing metrics from the current state of the art for different experimental settings, and pointers to further research that is needed to fill out our understanding of fair ranking measurement.

2 BACKGROUND

In this section we introduce several definitions and a brief summary of previous research concerning the fairness issue in IAS.

2.1 Algorithmic Fairness

Data-driven algorithmic systems, be they IAS or machine learning classification models, often reflect existing social biases in the data and context in which they are trained and evaluated into biases in their outcomes and effects [2]. Bias can appear throughout the design of such systems due to faulty data, flawed algorithms, biased evaluations, and other issues. In order to correct decision making reflections of systemic societal biases, it is essential to identify and measure bias. Fairness is hard to quantify; as an essentially contested social construct [29], there is not a single correct or objective definition of fairness.

Mitchell et al. [24] provide a survey of fairness definitions for classification models. One dimension along which fairness definitions divide is *individual* and *group* fairness. Individual fairness addresses the goal that similar individuals should (statistically) receive similar decisions, but crucially depends on a robust construct of similarity with respect to the task for which decisions are made, and there is currently no consensus in assignment of task-relevant similarity among individuals [6].

Group fairness, the focus on this work, aims to provide similar outcomes for members of different groups; this is often framed as ensuring a *protected group* is not treated unfairly with respect to a *dominant group*. Group membership is often defined by *sensitive attributes* such as race, gender, or ethnicity.

Two concepts that are particularly relevant to our work are *disparate treatment* and *disparate impact*; they are commonly used notions of unfairness to map with the concepts of *direct* and *indirect* discrimination. Disparate treatment occurs when different groups are intentionally treated differently, through either explicit use of sensitive attributes or other attributes designed to produce a discriminatory effect; disparate impact occurs when the system's effects are different for different groups regardless of intent [2]. *Statistical parity* — ensuring different groups receive favorable outcomes at comparable rates — is a common way of measuring disparate impact. Zafar et al. [36] introduced *disparate mistreatment* or *equalized odds* which defines the difference in error rates based on group association; a closely-related concept that informs some fair ranking designs is *equality of opportunity* [18], where qualified subjects should receive an equal probability of favorable outcome regardless of group membership.

2.2 Fairness in Information Access Systems

IAS introduce a further complication of being multi-sided environments, with different stakeholders having different fairness concerns [8]. Burke [7] distinguished between *provider fairness* and *consumer fairness* considering multiple stakeholders in recommender system frameworks. Provider fairness considers whether those who create the content a system recommends or retrieves are treated unfairly, on either a group or individual basis, while consumer fairness examines whether users experience the system unfairly. When studying provider fairness, the problem is further complicated by ranked outputs, particularly due to *position bias*:

[1] https://github.com/BoiseState/rank-fairness-metrics

users are more likely to see and engage with recommendations at the top of a list [35]. Slight changes to ranking may lead to large changes in the attention paid to a result and the economic return to its creator. Further, in any single ranking only one item may be placed at the top of the list, and will be the only item to accrue the benefits of first position regardless of the merit of the second item. Several metrics have been proposed to measure unfairness in ranking, variously taking into account user attention, exposure, and relevance of items. We survey these in more detail in Section 3, and refer the reader to the survey by Ekstrand et al. [13] for more exposition on fair information access in general. Beutel et al. [3] and Narasimhan et al. [25] take a different approach by defining fairness objectives over pairwise orderings instead of entire rankings, which is outside of the scope of this work.

In this paper, we focus on provider-side group-fairness of ranked outputs, adopting the common frame inspired by the United States anti-discrimination law of a "protected group": a class of people who share a trait upon which a recommendation or classification should not be discriminatory [33]. This includes discrimination on the basis of race, gender, religion, and similar traits. Some metrics are applicable to individual fairness; we note in their discussion.

3 FAIR RANKING METRICS

We begin by describing several fair ranking metrics, summarized in table 1, in a common framework and notation. This enables direct comparison of their designs and theoretical behavior, and facilitates easier implementation in IR experiments. In some cases, we assign new name for metrics based on their functionality, purpose, and comparability within our synthesis.

3.1 Problem Formulation

We consider an IR system that retrieves a ranked list L of n documents $d_1, d_2, \ldots, d_n \in D$ in response to requests (e.g. queries in a search system or users and/or contexts in a recommender system) $q_1, q_2, \ldots, q_n \in Q$ (notation summarized in table 2). Documents may have an associated request-specific relevance score $y(d|q)$, and the system may estimate this by a predictor $\hat{y}(d|q)$.

Providers are associated with one (or more) of g groups. We represent this by giving each document an alignment vector $\mathcal{G}(d) \in [0,1]^g$ (s.t. $\|\mathcal{G}(d)\|_1 = 1$) indicating its group association; generalizing from a categorical variable to a vector allows soft association (mixed or partial membership) or uncertainty about membership [27]. We generalize $\mathcal{G}(d)$ to a list function, with $\mathcal{G}(L)$ denoting an $n \times g$ alignment matrix whose rows correspond to the documents in L and columns are groups. In the case of definitively-known membership in a binomial pair of groups, $\mathcal{G}^+(L)$ denotes the set of documents in L in the "protected" group and $\mathcal{G}^-(L)$ the remaining documents (dominant group).

Our goal is to measure *exposure* (sometimes called *attention*) each document, content provider, or group receives, and assess the fairness of this distribution to ensure demographic or statistical parity (ensures comparable outcomes across groups) or *equality of opportunity* (ensures equal treatment based on merit or utility irrespective of the group membership). Accounting for the decreasing attention users are likely to pay to documents at deeper rank positions (*position bias*) requires a browsing model; some metrics build this implicitly into their structure, while others explicitly model it as a *position weight vector* $\mathbf{a}_L$ for L. Table 3 describes the various weighting schemes used by the metrics we survey. The resulting exposure is then sometimes compared with a *target distribution* $\hat{\mathbf{p}}$ that represents across groups. There are several ways of computing $\hat{\mathbf{p}}$, including strict group equality, an estimate of the population of actual or potential content providers, or the distribution among providers of relevant documents.

3.2 Statistical Parity in Single Rankings

We begin with metrics that assess the fairness of a single ranking and only measure exposure equity without considering relevance (that is, they target *statistical parity*). These metrics can be aggregated over the rankings produced by a system, e.g. by taking the mean, to produce an overall system fairness score.

The simplest way to measure the fairness of a single ranking is to measure the proportion of items in each group [14], but this does not account for position bias. Yang and Stoyanovich [34] propose a family of statistical parity measures that incorporate position bias by averaging parity over successive prefixes of the ranking; we call this the *prefix fairness* family (PreF_Δ). These metrics are optimized when the representation in each prefix matches the target $\hat{\mathbf{p}}$ as closely as possible, as measured by a distance function Δ; Yang and Stoyanovich used the full ranking's composition as $\hat{\mathbf{p}}$, and instantiate PreF_Δ with distance functions Δ_{ND}, Δ_{RD}, and Δ_{KD} (from Table 4) to yield different members of the family. The metric is defined as

$$\mathrm{PreF}_\Delta(L) = \frac{1}{Z} \sum_{i=10,20,30,\ldots}^{N} \frac{\Delta(L_{\leq i}, \hat{p})}{\log_2 i} \tag{1}$$

where normalizing scalar $Z = \max_{L'} \mathrm{PreF}'_\Delta(L', \hat{p})$ (taken over all L' with the same length and group composition as L, where PreF'_Δ is the prefix fairness function without the normalizer), scaling PreF_Δ to the range $[0,1]$ where 1 is maximum unfairness. Δ_{KL} has the advantage of allowing multinomial protected attributes and soft group association. PreF_Δ does not work when $\mathcal{G}^-(L) = \emptyset$, and Δ_{RD} does not work when $\mathcal{G}^-(L)$ is small. Z is also troublesome to compute with incomplete group membership data.

Zehlike et al. [37] propose a similarly-motivated group fairness constraint for a single list and fixed membership in binomial groups: L satisfies the FAIR constraint if for every prefix $L_{\leq k}$ with $1 \leq k \leq N$, the protected group is not statistically significantly under-represented. Unlike PreF_Δ, FAIR does not penalize over-representing the protected group. We convert this constraint into a metric by taking the average of the binomial probabilities:

$$\begin{aligned} \mathrm{FAIR}(L) &= \frac{1}{N} \sum_{k=1}^{N} P_{\mathrm{Binomial}} \left(m \leq |\mathcal{G}^+(L_{\leq k})| \quad \hat{p}, k \right) \qquad (2) \\ &= \frac{1}{N} \sum_{k=1}^{N} \sum_{j=1}^{|\mathcal{G}^+(L_{\leq k})|} \binom{k}{j} (\hat{p})^j (1-\hat{p})^{k-j} \end{aligned}$$

Sapiezynski et al. [27] provide a more general metric for single-list fairness by using an explicit (and configurable) position weight model instead of embedding the browsing model in the metric structure. Given an alignment matrix $\mathcal{G}(L)$ and suitably normalized position weight vector $\mathbf{a}_L$, $\epsilon_L = \mathcal{G}(L)^{\mathrm{T}} \mathbf{a}_L$ is a distribution that represents the cumulative exposure of the various groups in L.

Table 1: Summary of fair ranking metrics.

Metric(s)	Goal	Weighting	Target	Binomial?	Range	More Fair
PreF_Δ [34]	Each prefix representative of whole ranking	—	$\hat{p}$ from full ranking	Dep. on Δ^a	$[0, 1]$	0
AWRF_Δ [27]	Weighted representation matches population	Geometric	configured $\hat{p}$	Dep. on Δ	$[0, 1]$	0
FAIR [37]	Each prefix matches target distribution	—	binomial $\hat{p}$	Yes	$[0, 1]$	0
logDP [30]	Exposure equal across groups	Logarithmic	equality	Yes	$(-\infty, \infty)$	0
logEUR [30]	Exposure proportional to relevance	Logarithmic	$\propto$ utility	Yes	$(-\infty, \infty)$	0
logRUR [30]	Discounted gain proportional to relevance	Logarithmic	$\propto$ disc. utility	Yes	$(-\infty, \infty)$	0
IAA [5]	Exposure proportional to predicted relevance	Geometric	$\propto$ est. utility	No	$[0, \infty)$	0
EEL, EER [11]	Exposure matches ideal (from relevance)	Cascade, RBP.	f(utility)	No	$[0, \infty)$	EEL 0, EER >
EED [11]	Exposure well-distributed	Cascade^b, RBP.	equality	No	$[0, \infty)$	0

[a] Δ_{RD} and Δ_{RD} both require binomial protected group attributes, but Δ_{KL} generalizes.
[b] Cascade weighting also incorporates relevance into exposure, even if exposure is not compared to relevance.

Table 2: Summary of notation.

$d \in D$	document or item
$q \in Q$	request (query or user)
L	ranked list of N documents from D
$L^{-1}(i)$	the document in position i of list L
$L(d)$	rank of document d in L
$L_{\leq k}$	prefix of L of length k
$y(d \mid q)$	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 ($\mathcal{G}(L)^T \mathbf{a}_L$)

The resulting unfairness metric, which we call *Attention-Weighted Rank Fairness* (AWRF_Δ), is the difference between this exposure distribution and the population estimator:

$$\mathrm{AWRF}_\Delta(L) = \Delta(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}}) \tag{3}$$

AWRF_Δ allows soft association and multinomial protected attributes. The distance function in Table 4 depends on application context; for assessing a particular protected class representation, difference in probability is suitable distance.

3.3 Statistical Parity in Multiple Rankings

In many cases, fair exposure cannot be achieved in a single ranking, because the attention paid to rank positions often decreases more steeply than the utility (relevance) of documents [5, 11]. One solution is to measure fairness over sequences or distributions of rankings so providers have comparable opportunity to be exposed in at least some sessions or responses. This approach can be modeled as a request-dependent distribution (or *policy*) $\pi(L|q)$ over rankings [11, 30]. We extend this to include a distribution over requests $\rho(q)$, so a sequence of rankings $L_1, L_2, \ldots, L_{\tilde{n}}$ [5] is a series of draws from the distribution $\rho(q)\pi(L|q)$. The group exposure within a single ranking from Eq. 3, $\boldsymbol{\epsilon}_L = \mathcal{G}(L)^T \mathbf{a}_L$, is the fundamental building block of these metrics, along with its expected value:

$$\boldsymbol{\epsilon}(q) = \mathrm{E}_\pi[\boldsymbol{\epsilon}_L] = \sum_L \pi(L \mid q) \boldsymbol{\epsilon}_L$$

$$\boldsymbol{\epsilon}_\pi = \mathrm{E}_{\pi\rho}[\boldsymbol{\epsilon}_L] = \sum_q \rho(q) \boldsymbol{\epsilon}(q)$$

Singh and Joachims [30] and Diaz et al. [11] each propose metrics for measuring statistical parity over ranking policies. Neither metric incorporates a target distribution; they are optimal when all groups are equally exposed. Demographic parity [DP, 30] measures the difference in exposure between two groups:[2]

$$\mathrm{DP} = \boldsymbol{\epsilon}_\pi(\mathcal{G}^+) / \boldsymbol{\epsilon}_\pi(\mathcal{G}^-) \tag{4}$$

Expected exposure disparity [EED, 11] ensures well-distributed exposure by measuring the inequality in exposure distribution across groups with the L_2 norm:

$$\mathrm{EED} = \|\boldsymbol{\epsilon}_\pi\|_2^2 \tag{5}$$

3.4 Equal Opportunity in Multiple Rankings

So far, none of the metrics we have discussed account for the utility of the ranked results — rankings do well by exposing providers regardless of the utility of their items. The intuition behind incorporating utility, articulated independently by Singh and Joachims [30] and Biega et al. [5], is that exposure should be proportional to relevance: if an item or a group contributes 10% of the relevance to a request (user and/or query), it should receive approximately 10% of the exposure. This is a ranked analog of the *equality of opportunity* construct from fair classification [18]: outcome is conditionally independent of group given utility.

To measure deviation from this goal, Singh and Joachims [30] propose two metrics. The *exposed utility ratio* (EUR)[3], measures deviation from the goal that each group's exposure is proportional

[2] The original paper presented a constraint, not a metric, for demographic parity; we have implemented it as a ratio to be consistent with the other metrics.

[3] Singh and Joachims [30] used the terms "disparate treatment ratio" and "disparate impact ratio" for EUR and RUR, respectively, but this terminology is not consistent with the use of these terms in the broader algorithmic fairness literature as we understand it. Exposure the system gives to providers is an impact, not a treatment. We have changed the names to hopefully reduce confusion going forward.

Table 3: Weighting models for computing $\mathrm{a}_L(d)$ with default parameter values.

Metric	Model	Formula	Parameters
AWRF, IAA	Geometric	$\gamma(1-\gamma)^{L(d)-1}$	Stopping probability γ
logDP, logEUR, logRUR	Logarithmic	$1/\log_2 \max\{L(d), 2\}$	—
EER, EED, EEL	RBP	$\gamma^{L(d)}$	Continuation probability (patience) γ
EEL, EED, EER	Cascade	$\gamma^{L(d)-1} \prod_{j\in[0,L(d))} \left[1-\phi\left(y(L^{-1}(j) \mid y)\right)\right]$	Patience γ, stopping probability function ϕ

Table 4: Distance functions for comparing distributions.

Distance Function	$\hat{\mathbf{p}}^a$	Formula
$\Delta_{\mathrm{ND}}(L, \hat{\mathbf{p}})$	Binomial	$\frac{\lvert\mathcal{G}^+(L)\rvert}{N} - \hat{\mathbf{p}}$
$\Delta_{\mathrm{RD}}(L, \hat{\mathbf{p}})$	Binomial	$\frac{\lvert\mathcal{G}^+(L)\rvert}{\lvert\mathcal{G}^-(L)\rvert} - \frac{\hat{\mathbf{p}}}{1-\hat{\mathbf{p}}}$
$\Delta_{\mathrm{KL}}(L, \hat{\mathbf{p}})$	Multinomial	$D_{\mathrm{KL}}(\hat{\mathbf{p}}(L) \Vert \hat{\mathbf{p}})^b$
$\Delta_{\mathrm{AD}}(\boldsymbol{\epsilon}_L, \hat{\mathbf{p}})$	Binomial	$\lvert\frac{\lvert\mathcal{G}^+(L)\rvert}{N} - \hat{\mathbf{p}}\rvert$

[a] Binomial $\hat{\mathbf{p}}$ is a scalar probability of the protected group.
[b] K-L divergence; $\hat{\mathbf{p}}(L)$ is the probability distribution of groups in L.

to its contributed utility (measured by $\mathrm{Y}(\mathcal{G}) = \mathrm{E}_\rho[\frac{1}{g}\sum_{i\in g} y(d|q)]$):

$$\mathrm{EUR} = \frac{\boldsymbol{\epsilon}_\pi(\mathcal{G}^+)/\mathrm{Y}(\mathcal{G}^+)}{\boldsymbol{\epsilon}_\pi(\mathcal{G}^-)/\mathrm{Y}(\mathcal{G}^-)} \tag{6}$$

The *realized utility ratio* (RUR) incorporates utility into both numerators and denominators by measuring whether the *discounted* utility contributed by each group ($\Gamma(\mathcal{G}) = \sum_{d\in\mathcal{G}} \mathrm{E}_{\pi\rho}[\mathbf{a}_L(d)y(d|q)]$) is proportional to its total utility:

$$\mathrm{RUR} = \frac{\Gamma(\mathcal{G}^+)/\mathrm{Y}(\mathcal{G}^+)}{\Gamma(\mathcal{G}^-)/\mathrm{Y}(\mathcal{G}^-)} \tag{7}$$

As they are based on ratios between group metrics, EUR and RUR do not support multinomial protected groups or soft association.

Biega et al. [5] present the *amortized attention* construct to measure exposure over the sequence of rankings. This compares rank exposure with expected utility $\hat{\mathrm{Y}}$ (computed with system-predicted utility $\hat{y}(d|q)$) instead of ground truth relevance assessments $y(d|q)$), measuring whether the system allocates exposure proportional to the utility it estimates items to have. Deviations from this goal are measured by taking the L_1 norm of the group exposure-utility differences, yielding the *Inequity of Amortized Attention* (IAA) metric:

$$\mathrm{IAA} = \|\boldsymbol{\epsilon} - \hat{\mathrm{Y}}\|_1 \tag{8}$$

Diaz et al. [11] build on this to integrate relevance in a different way. Rather than relate exposure directly to relevance, they use relevance to derive *target exposure* based on an ideal policy τ that assigns equal probability to all rankings that place items in non-decreasing order of relevance and 0 (or miniscule) probability to all other rankings. The target exposure $\boldsymbol{\epsilon}^*$ is the expected exposure under the ideal policy ($\boldsymbol{\epsilon}^* = \mathrm{E}_{\tau\rho}[\boldsymbol{\epsilon}_L]$). They take the squared Euclidean distance between system expected exposure and target exposure, yielding the *Expected Exposure Loss*:

$$\mathrm{EEL} = \|\boldsymbol{\epsilon}_\pi - \boldsymbol{\epsilon}^*\|_2^2 \tag{9}$$

$$= \|\epsilon_\pi\|_2^2 - 2\boldsymbol{\epsilon}_\pi^\top\boldsymbol{\epsilon}^* + \|\boldsymbol{\epsilon}^*\|_2^2 \tag{10}$$

The decomposition in Eq. 10 yields *expected exposure relevance* $\mathrm{EER} = 2\boldsymbol{\epsilon}_\pi^\top\boldsymbol{\epsilon}^*$ (measuring the alignment of exposure and relevance, higher values represent better alignment) along with EED. Neither IAA nor the EE metrics distinguish between group over- or under-exposure; for both, 0 is perfectly fair and larger values are unfair, with no preferential treatment given to a protected group.

The common thread between these metrics, articulated by Diaz et al. [11], is that for a fixed information need, differences in exposure between items with the same relevance grade results in unjustifiably unfair outcomes. Relating exposure to relevance sets the goal that items of comparable relevance should have comparable opportunity to be exposed, as measured by expected or amortized exposure over repeated rankings.

3.5 Assessing Metric Design

Rendering metrics in a common notation shows that the metrics are quite similar in their basic concepts. The fundamental construct — weighted exposure — is the same across most metrics, and they differ primarily in how they relate exposure to relevance and how they aggregate and compare exposure distributions. The following questions help identify more precisely what their salient differences are and how those may relate to particular IAS applications and experimental settings.

Does the metric incorporate relevance? EEL, EER, EUR, RUR, and IAA directly incorporate relevance into metric; others strictly measure statistical parity. It is desired depending on the precise task and evaluation goal. Statistical parity metrics are useful for measuring relative fairness of rankings already optimized for utility, particularly when there is no relevance information available or the relevant sets for a query are large. They can also be used to detect discrepancies that may indicate unfairness in relevance data (if relevance data is unfair, for example by systematically underestimating the relevance of a group's documents, a metric that relates exposure to relevance will use the unfair relevance to justify unfair disparities in exposure). However, using such metrics in isolation for evaluation or optimization may reduce ranking quality.

How does it handle missing data? Real-world data sets are often incomplete, missing relevance and/or group labels for many documents. Metrics that are less sensitive to that problem will be easier to apply in such cases. Missing relevance data affects EUR, RUR, EER, and EEL like it does classical IR evaluation metrics such as nDCG; the straightforward but biased approach is to treat items with unknown relevance as irrelevant ($y = 0$). IAA's use of system-estimated relevance allows it to sidestep this problem.

Missing group labels require different handling. For many metrics we can include unlabeled items when computing attention weights but exclude them from further analysis, or treat "unknown" as an additional group identity. Unknown data is a more significant problem for PreF_Δ family because it treats a list with fewer than 10 known-group items as maximally fair, and the straightforward way of computing Z — make the ranking maximally unfair by putting all protected items last — does not work in the face of missing data.

How does it respond to edge cases? Realistic IR experiments bring a number of important edge cases, such as groups with no items relevant to or retrieved for a request. Ratio-based metrics and distance functions are particularly vulnerable to these problems; the EUR metric and the Δ_{RD} distance function, for example, approach infinity as the number of non-protected-group items retrieved goes to zero. RUR is even more brittle, as it requires nonzero relevance from retrieved non-protected-group items to avoid infinity, and both it and EUR can be infinite or undefined if the set of relevant items from either group is zero.

Reformulation of DP*,* EUR *and* RUR*:* Since these three metrics are ratios, their maximally fair point is 1, with a nonlinear relationship between values favoring and disfavoring the protected group, hindering interpretability; further, they approach ∞ if the dominant group exposure is close to 0. To improve interpretability, we take the logs of these ratios, so 0 is fair and distance is symmetric in either direction; and we address the empty-group problem by adding a small damping constant to both sides of the ratio. This yields the following reformulation for DP:

$$\mathrm{logDP} = \log\left(\epsilon(G^{+}) + 10^{-6}\right) - \log\left(\epsilon(G^{-}) + 10^{-6}\right)$$

logEUR and logRUR are defined equivalently. As log ratios, values greater than 0 indicate a bias in favor of the protected group.

What is the target? PreF_{Δ}, FAIR, AWRF_{Δ}, EEL, and EER provide flexibility in determining how the (un)fairness of exposure is ultimately assessed through selection of the target distribution, while targets are implicitly baked in to the structure of others. This configurability is useful because it allows the metric to be adapted to the fairness requirements of a particular task, although it can impair comparability between experiments.

How does the metric compare the system with the target? Some metrics (AWRF_{Δ} and PreF_{Δ}) use an explicit distance function to compare distributions, while others use ratios of specific proportions or norms of differences in distributions. Norms and selected distance functions (such as Δ_{KL}) can accommodate soft association, while ratios and distance functions based on binomial probabilities require definitive membership in binomial groups. They can be adapted to some multi-group situations if only one group's exposure needs to be considered.

What user model does it use? Most metrics allow different position weighting strategies to be selected, both in its structure and its parameters. This configurability allows the metric to be adapted to specific application but introduces potential sensitivity to choices of weight functions and parameter values. PreF_{Δ} and FAIR are not configurable, as position weighting is built-in.

4 EXPERIMENTAL SETUP

We now turn from our analytical treatment of the metrics to an empirical comparison, using each of them (except for PreF_{Δ}, due to its difficulties with missing labels and soft membership) in real world IAS experiments for three tasks across two problem settings:

(1) Personalized book recommendations, measuring fairness with regards to author gender.

(2) Scholarly article retrieval (both retrieval and re-ranking of short candidate sets) based on queries, measuring fairness with regard to the economic development of the author's country (as a proxy for the research resources available).

Table 5: Summary of experiment data.

		Fair TREC 2020	
	GoodReads	Rerank	Retrieval
Systems	4	23	5
Requests	5000	195	189
Items	23,60,655	2112	2112
$\lvert\mathcal{G}^{+}\rvert$	1,90,711	294	294
$\lvert\mathcal{G}^{-}\rvert$	21,17,451	1632	1632

We further carry out a sensitivity analysis to understand how experimental outcomes change in response to design decisions and parameter values in the metrics.

This section describes the experimental setup itself, and the considerations we had to make when adapting the metrics in this setting. Table 5 shows summary statistics for each dataset.

4.1 Recommendation (GoodReads)

For our recommendation experiments, we used the *GoodReads* data [32] in an experiment adapted from that of Ekstrand and Kluver [14], using Version 2.0 of the PIReT Book Data Tools[4] to prepare data sources. We use LensKit for Python [12] to train recommendation models on implicit feedback data from GoodReads, with a positive user-book interaction if the user ever added the book to a shelf. We sampled 5000 users for our experiment, each of which had at least 10 book interactions, holding out 5 interactions per user as test data for assessing relevance in the resulting recommendations. We used four collaborative filtering (CF) algorithms: user-based CF (UU [19]), item based CF (II [10]), matrix factorization (WRLS [31]), and Bayesian Personalized Ranking (BPR [26]), using configurations and hyperparameter tunings from Ekstrand and Kluver [14], to generate a single list of 100 recommendations for each user.

We measured the fairness of each recommendation list with respect to the gender of the book's author, extracted from Virtual Internet Authority File (VIAF)[5] (as described by Ekstrand and Kluver [14]). Group membership in this data is binary but incomplete, so we considered female authors to be the protected group G^{+} and male authors G^{-} for all two-group metrics (unknown-author books are therefore ignored). For AWRF_{Δ}, we used Δ_{AD} (following the original presentation [27]), and the distribution of male and female authors among the set of books in the data set as the population estimator. For IAA and the EE metrics, we treat unknown gender as a third author group.

4.2 Search (FairTREC)

For search experiments, we used submitted runs and evaluations from the *TREC Fair Ranking Track* 2020 [4]. These runs covered two tasks (re-ranking and full retrieval). We considered each submitted

[4] https://bookdata.piret.info

[5] http://viaf.org/viaf/data/

run as an individual system and used the given sequences of rankings for each system. For the re-ranking task, we only used one run from each participating team. The details about the systems can be found in participants' notebook papers [15, 21, 23, 28]

Unlike GoodReads, in the FairTREC data, each document has a soft association with the economic development level of its author(s), and thus we could not implement logDP, logEUR, and logRUR. Additionally, IAA uses system predicted relevance as ground truth, which makes it inapplicable in TREC setup (because systems do not provide scores for all items)

Table 6: Default configuration for metrics

Metrics	Weighting	Stop %prob.	Patience
AWRF$_\Delta$	Geometric	0.5	—
logDP, logEUR, logRUR	Logarithmic	—	—
IAA	Geometric	—	0.5
EEL, EER, EED	RBP	—	0.5
EEL, EER, EED	Cascade	0.5	0.5

5 EMPIRICAL RESULTS

We now present the results of our experiment, using both the metrics in their default configurations and conducting a sensitivity analysis with respect to weighting methods and parameters.

5.1 Direct Comparison

We begin by directly comparing the metrics with default parameter settings from their original papers to see how they assess each system in our experiments. Table 3 shows the default configuration of the metrics. This comparison allows us to get a first view of the differences in results using each metric as originally presented, with minimal adjustments for practical implementation (see Section 3).

Figure 1 shows the metric results from our experiments. From this we observe two things:

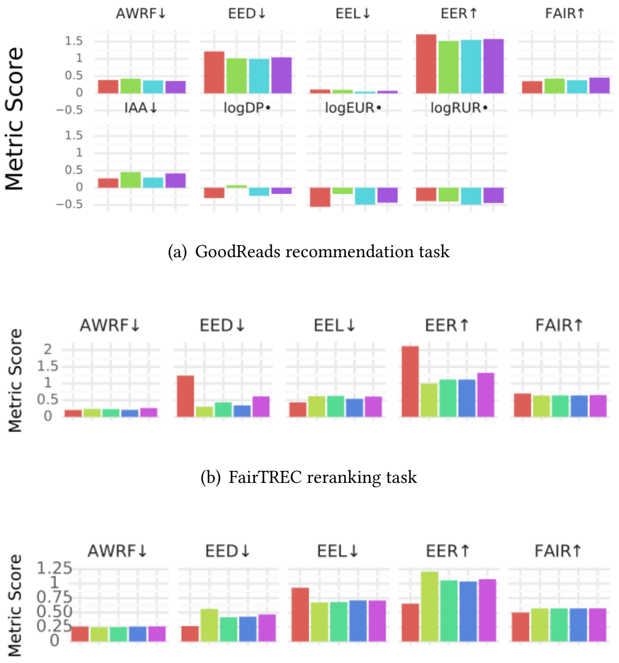
- Metrics frequently disagree on system orderings.
- Metrics that agree in one experiment don't necessarily agree on others. The most consistently-agreeing pair is FAIR and AWRF$_\Delta$, the two single-list metrics we study.

5.2 Sensitivity Analysis

Section 3 demonstrates that the fair ranking metrics often incorporate several design choices. However, this does not tell us how much difference these choices make in practice; if a metric is highly sensitivity towards design choices, it is more difficult to make correct configuration decisions (particularly in the absence of external guidance for those choices), increasing the complexity of applying it and the likelihood of error. To further analyze the applicability and sensitivity of these metrics, we need to know to what extent these metrics are dependant on their decision choices. We now turn to understanding the impact of design decisions and parameter settings *within* each metric.

As we noted in Section 3, the exposure-based metrics and AWRF$_\Delta$ combine position weights and relevance in various ways; each was presented with particular position weighting strategy, but could be applied to any other. Further, most weighting strategies have parameters that affect the strength of the discounting. We test the sensitivity of metrics and conclusions drawn from them to the choice of ranked-list size, position weight formula, patience parameter, and stopping probability.

Figure 1: Fairness metrics *GoodReads* and *FairTREC* datasets using their original configurations. Arrow indicates direction of maximal fairness; · means 0 is fair. System identities are not relevant to our results.



(a) GoodReads recommendation task

(b) FairTREC reranking task

(c) FairTREC full retrieval task

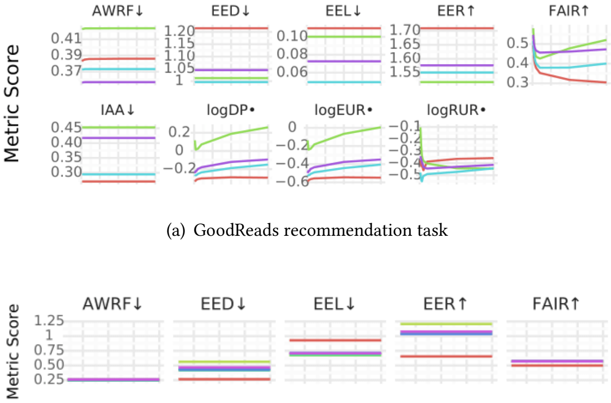
5.2.1 Size of Ranked List. To observe the sensitivity towards ranked list size we apply the metrics lists of varying sizes (10–1000 for GoodReads recommendation, and 10–100 for FairTREC full retrieval). Fig. 2 shows the outcome of fairness metrics with the change of ranking length. We observe that:

- Changing ranked list size had no effect on any metric applied to FairTREC.
- AWRF$_\Delta$, *IAA*, *EEL*, *EED*, and *EER* are mostly stable as the list length changes in the GoodReads recommendation experiment; they show slight changes through length 50, but without affecting system ordering, and then stabilize.
- *logDP, logEUR* and *FAIR* (on GoodReads) change notably, including reordering algorithms, as the list size changes.

Sensitivity towards ranked-list size of ratio-based metrics and FAIR in recommendation task indicates to the need of studying metric dependency on relevance and group information availability.

5.2.2 Weighting Strategy. For position-weighted metrics, we applied each metric to all four position weight models: *rbp*, *cascade*, *geometric*, and *logarithmic* (summarized in Table 3). Figure 3 shows

Figure 2: Metric results with the change of ranked-list size.



(a) GoodReads recommendation task

(b) FairTREC full retrieval task

the outcome of fairness metrics with the change of position weighting strategy. We use a continuation probability of 0.5 for the patience parameter and a stopping probability of 0.5. From these results, we observe:

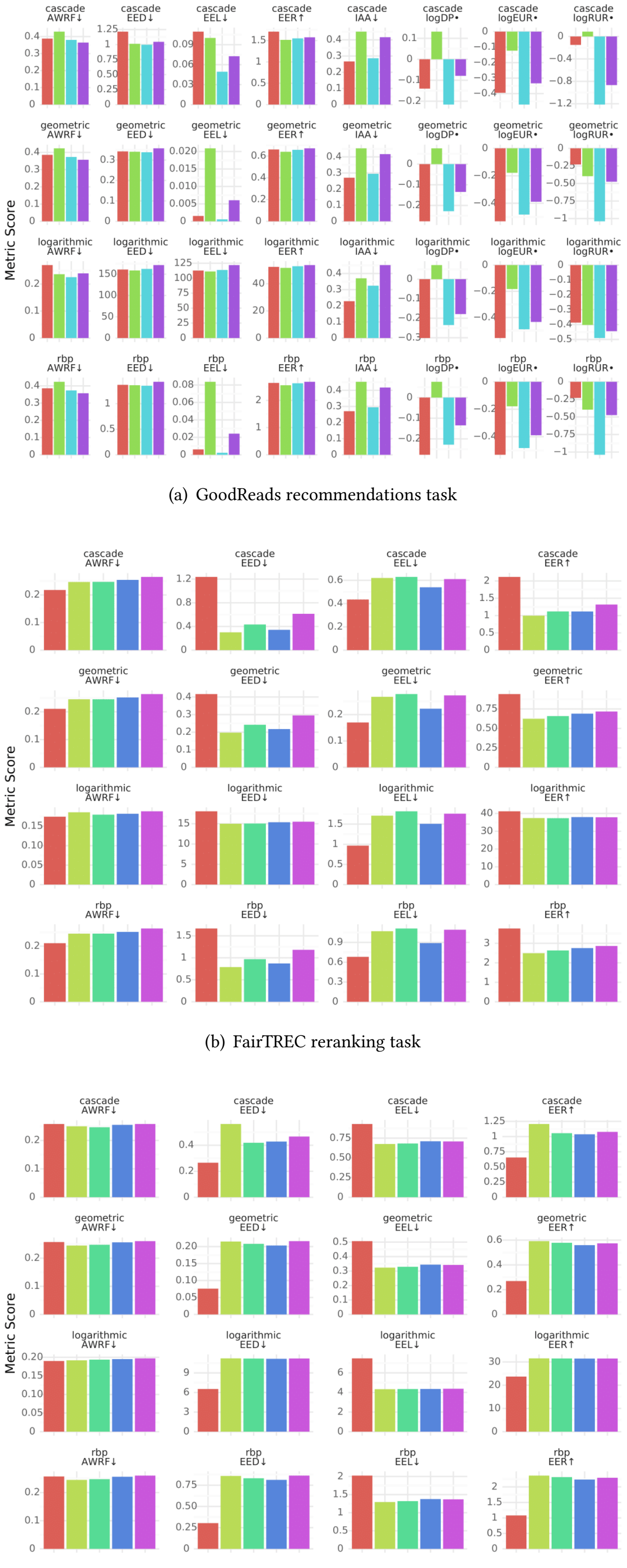
- For GoodReads recommendation task, logDP, logEUR, and AWRF$_\Delta$, systems show differences with the change of weighting strategy, whereas for IAA, EER and EED, algorithms remain stable and did not show much disagreement across different weighting strategies. logRUR and EEL show extreme sensitivity towards the change of weighting model.
- In FairTREC reranking (Fig. 4(b)), systems show small differences but generally maintain system orderings across weighting models. We observe a few changes in order (e.g. AWRF$_\Delta$ from cascade to logarithmic) but these are between systems already very close.
- In FairTREC full retrieval (Fig. 4(c)), systems are generally stable across position models.

From the analysis, we observe that browsing models can have effects over some metrics' behavior to some extent, specially on EEL and logRUR. However, this analysis does not let us conclude that these metrics which showed stability over various weighting strategies will act uninfluenced with the change of parameters in weighting strategies. For further investigation, we measure the metrics by changing the parameter values.

5.2.3 Patience Parameter. Figure 4 presents the response of the metrics across patience parameter changes for the *rbp* and *cascade* weightings where we can see the following patterns:

- AWRF$_\Delta$, EEL, EER, and EED show sensitivity towards the patience parameter following the same pattern in all three tasks (full retrieval, reranking, recommendations)
- In EEL, EED, and EER, systems show mild separation with each other following the same pattern across weighting strategies.
- On GoodReads recommendation tasks, logDP, logEUR, and IAA show substantial separation between systems; they also

Figure 3: Metric results with the change of weighting strategy.

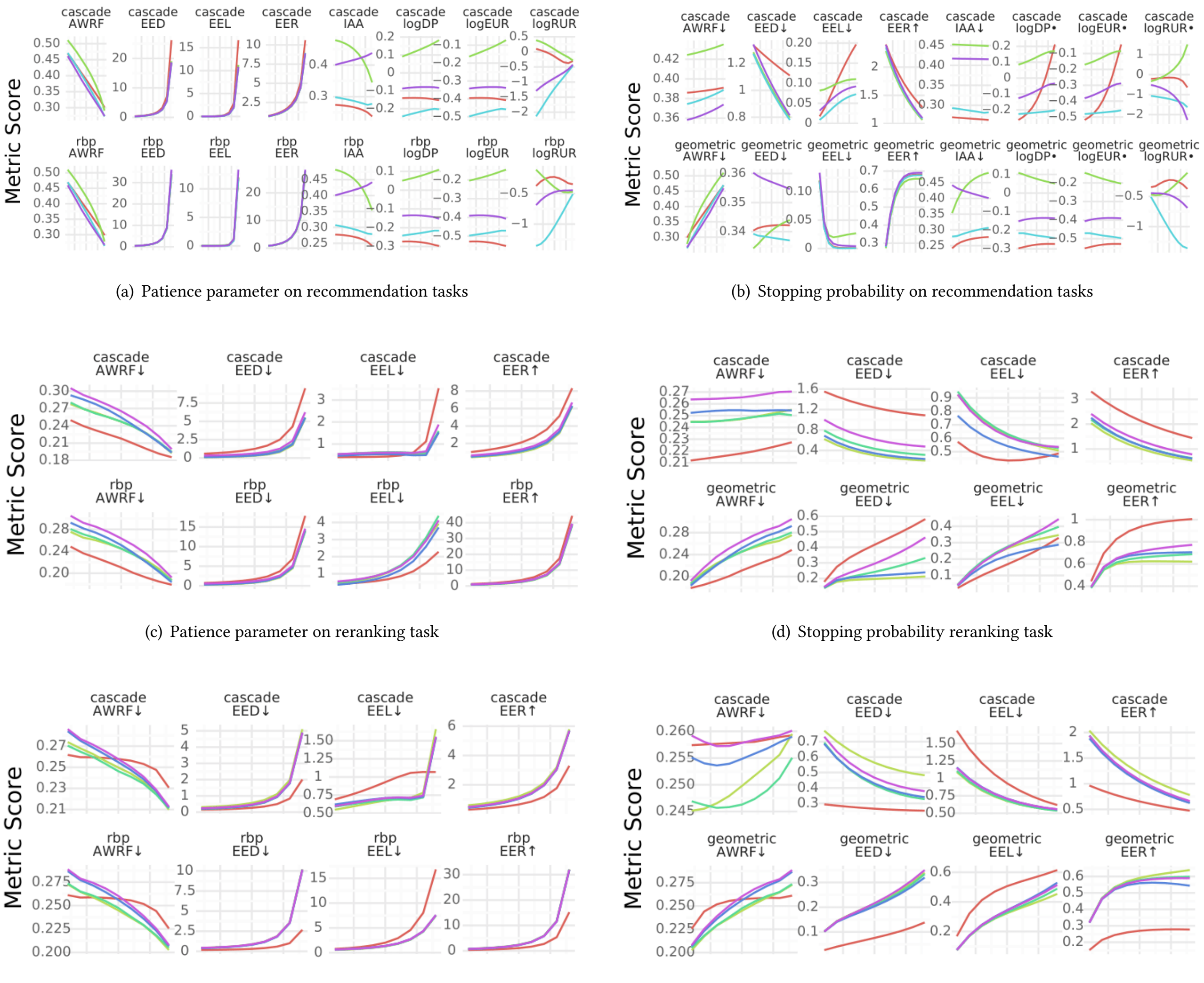


(a) GoodReads recommendations task

(b) FairTREC reranking task

(c) FairTREC full retrieval task

Figure 4: Metric results with the change of patience parameter and stopping probability.



(a) Patience parameter on recommendation tasks

(b) Stopping probability on recommendation tasks

(c) Patience parameter on reranking task

(d) Stopping probability reranking task

(e) Patience parameter on full retrieval task

(f) Stopping probability on full retrieval task

preserve system order as the parameter changed but the differences between systems shifted. The systems follow a similar pattern across weighting strategies.

- logRUR is extremely sensitive to patience parameter changes.

5.2.4 Stopping Probability. Figure 4 shows the outcome of fairness metrics on the generated recommendations for *geometric* and *cascade* position weight models and the sensitivity towards the change of stopping probability. We have made the following observations from the charts:

- In FairTREC full retrieval and reranking tasks, metric results change with stopping probability. However, the systems did not vary in the changing pattern significantly.
- On the GoodReads recommendations task (figure 5(b)), IAA, logDP and logEUR shows sensitivity with the change of stopping probability and the sensitivity is notable in the cascade weighting strategy.
- In all three tasks, systems show complete inversion across weighting strategies for EEL, EED, and EER. In EEL, the pattern of sensitivity towards stopping parameter is different between recommendation and ad-hoc tasks.
- logRUR is extremely sensitive to patience parameter changes.

Almost all metrics show sensitivity towards parameter value changes, which imply the necessity of identifying optimal parameter settings while implementing these metrics.

Overall, we observe that metrics do vary in their responses with the change of design choices, however, IAA, EER, EED and AWRF$_\Delta$ showed the most stability.

6 DISCUSSION AND RECOMMENDATIONS

We started this project with three goals:

1. Identify requirements to implement the fair ranking metrics in actual search and recommendation frameworks.
2. Identify similarities and both analytical and empirical differences among metrics to inform the metric selection process.
3. Identify the observable effects of different changes in the metric design or configuration.

Our analysis provides significantly more in-depth knowledge about the fairness goals, requirements, implementations, and effect of design decisions. In summary, our key findings are the following:

- Many metrics are remarkably similar in their underlying concept of fairness.
- Metric implementation highly relies on crucial factors such as group size, ranked list size, item relevance information, and group membership.
- Certain design choices can make metrics vulnerable to edge cases. For example, ratio-based metrics have difficulties with empty groups and zero values, such as a ranking that has no retrieved items from one of the groups.
- Despite having similar fairness goals, these metrics can differ in their sensitivity towards external factors.

This still leaves the question, however, of what we should do in the present to measure (un)fairness in ranking from real information retrieval system datasets using a fair ranking metric. We propose to use metrics compatible with the following criteria:

- Allow multinomial protected attributes. Such metrics are applicable to a wider range of fairness settings, and choosing one means that the metric is not a reason to use a binary simplification of a multinomial attribute, such as gender.
- Allow soft group association (mixed or partial membership).
- Be stable with respect to design choices.

This last point is to support ease of use; if a metric is highly sensitive to design choices such as the attention weighting model, then its validity depends more strongly on the correctness of those choices. While a metric's validity with respect to the fairness objective in a particular application setting is the most important factor, given two comparably-appropriate metrics we would prefer one that is more robust to misspecification in its configuration. Based on these requirements, combined with our observations in sections 3 and 5, we make recommendations for different measurement goals and context based on the current state of the art and knowledge about fair ranking metrics:

Single Rankings. All single-ranking metrics we considered are statistical parity metrics — they do not incorporate relevance. From our analysis, AWRF_{Δ} seems the most generally useful, because it supports multinomial protected attributes with soft assignment, and is adaptable to multiple attention models, target distributions, and difference functions. We are not yet able to make concrete recommendations for the choice of a difference function.

Demographic Parity in Sequences. logDP and EED measure statistical parity on sequences of rankings. **EED** seems more generally useful because of its support for multinomial groups with soft membership, and was relatively robust with respect to design choices.

Equal Opportunity in Sequences. The logEUR, logRUR, IAA, EER, and EEL metrics use relevance to measure (un)fairness in sequence of ranking, aiming at some version of equality of opportunity. We currently recommend using **EER** and **EEL** because of their support for multinomial groups with soft assignment, and comparative robustness. IAA shows comparable stability and can be adapted to multinomial groups and soft assignment; exploring that possibility is future work. In each context, position weighting model should be chosen based on user behavior in the expected context of use.

7 CONCLUSION AND FUTURE DIRECTION

This paper presents a comparative analysis among several fairness metrics recently introduced to measure fair ranking. We discuss the metric formulations and implications in an integrated notation and presented the first (to our knowledge) empirical comparison of fair ranking metrics for recommendation and search systems in common data sets and fairness goals. We hope this comprehensive presentation and comparison among metrics will help future researchers and practitioners to make more informed decisions about metric choice and configuration.

Our findings from this empirical analysis point to several directions for future research. Further work is needed on the limitations we observe from implementing the metrics in real data: implications and corrective methods for missing or sparse relevance information of items and missing [20], ambiguous, or multiple group associations [17] are not yet well-understood. Moreover, the instabilities we observe in our sensitivity analysis points to the need to work towards designing robust and efficient fair ranking metrics and develop a body of research that can lend external support for choosing where in the design space best meets a particular fairness goal. We also expect simulation studies will yield a much deeper insight into the differences we observed applying metrics across different tasks and datasets, understanding more thoroughly the impact of factors like relevant set size, soft association, and missing relevance information, among others.

Significant progress has been made in the last 2–3 years on measuring the fairness of rankings, but more work is needed in order to understand how best to design and apply these metrics.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS 17-51278. We also thank our colleagues in the People and Information Research Team (PIReT) for their support, and particularly Connor Wood for his assistance in implementing the metrics.

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity Bias in Ranking and Recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 529–530. https://doi.org/10.1145/3306618.3314309

[2] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *Calif. L. Rev.* 104 (2016), 671.

[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, Fernando Diaz, Michael D Ekstrand, and Sebastian Kohlmeier. 2020. Overview of the Trec 2019 Fair Ranking Track. *arXiv preprint arXiv:2003.11650* (2020).

[5] 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

[6] Reuben Binns. 2020. On the Apparent Conflict Between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524. https://doi.org/10.1145/3351095.3372864

[7] Robin Burke. 2017. Multisided Fairness for Recommendation. (July 2017). arXiv:1707.00093 [cs.CY] http://arxiv.org/abs/1707.00093

[8] Robin D Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. 2016. Towards Multi-Stakeholder Utility Evaluation of Recommender Systems.. In *ACM UMAP Conference on User Modeling, Adaptation and Personalization (Extended Proceedings)*.

[9] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336. https://doi.org/10.1145/290941.291025

[10] 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

[11] 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

[12] 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

[13] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness and Discrimination in Information Access Systems. *Foundations and Trends in Information Retrieval* (2022). https://arxiv.org/abs/2105.05779

[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] Yunhe Feng, Daniel Saelid, Ke Li, Ruoyuan Gao, and Chirag Shah. 2020. University of Washington at TREC 2020 fairness ranking track. *arXiv preprint arXiv:2011.02066* (2020).

[16] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[17] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. *When Fair Ranking Meets Uncertain Inference*. Association for Computing Machinery, New York, NY, USA, 1033–1043. https://doi.org/10.1145/3404835.3462850

[18] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv preprint arXiv:1610.02413* (2016).

[19] 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.

[20] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1065–1075. https://doi.org/10.1145/3442381.3450080

[21] Till Kletti and Jean-Michel Renders. 2020. Naver Labs Europe at TREC 2020 Fair Ranking Track. (2020).

[22] 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)*.

[23] Graham McDonald and Iadh Ounis. 2020. University of Glasgow Terrier Team at the TREC 2020 Fair Ranking Track. In *The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings*, Vol. 1266.

[24] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (2021).

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

[26] 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.

[27] 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

[28] Mahmoud F Sayed and Douglas W Oard. 2020. The University of Maryland at the TREC 2020 Fair Ranking Track. (2020).

[29] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[30] 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

[31] 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

[32] 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

[33] A Xiang and I Raji. 2019. On the Legal Compatibility of Fairness Definitions. In *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. https://arxiv.org/abs/1912.00761

[34] 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.

[35] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 1011–1018. https://doi.org/10.1145/1772690.1772793

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[37] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938

[38] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. *arXiv preprint arXiv:2103.14000* (2021).