Pessimistic Evaluation

Fernando Diaz
Google
Pittsburgh, PA, United States
diazf@acm.org

Abstract

Traditional evaluation of information access systems has focused primarily on average utility across a set of information needs (information retrieval) or users (recommender systems). In this work, we argue that evaluating only with average metric measurements assumes utilitarian values not aligned with traditions of information access based on equal access. We advocate for pessimistic evaluation of information access systems focusing on worst case utility. These methods are (i) grounded in ethical and pragmatic concepts, (ii) theoretically complementary to existing robustness and fairness methods, and (iii) empirically validated across a set of retrieval and recommendation tasks. These results suggest that pessimistic evaluation should be included in existing experimentation processes to better understand the behavior of systems, especially when concerned with principles of social good.

CCS Concepts

• **Information systems → Evaluation of retrieval results**.

Keywords

Evaluation, Information Retrieval, Recommender Systems, Fairness

ACM Reference Format:
Fernando Diaz. 2024. Pessimistic Evaluation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24), December 9–12, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3673791.3698428

1 Introduction

Evaluating information access systems that support large populations of users remains a fundamental area of research. *Individual-level evaluation*, the measurement of the utility that an individual receives from the system in a specific context, has been the focus of work in metric design familiar to the information retrieval community. *Population-level evaluation* refers to making judgments about a system based on how individual-level utility is distributed across a group or population of users; the most common approach to population-level evaluation is to use the arithmetic mean utility as the measure of system quality. While individual-level evaluation metrics can often be empirically validated against human preferences, population-level evaluation methods often do not have quantifiable signals that can be used for validation. Instead, population-level evaluation methods often need to be more rigorously validated for their alignment with normative values about how utility should be distributed across a population. This is consistent with research in the recommender systems community seeking to incorporate of normative values into system evaluation and design [14, 15, 39–41]. Developing evaluation methods with rigorous conceptual and theoretical foundations is fundamental to translational work between system policy and development.

While the arithmetic mean dominates experimental practice, system design objectives have led to the development of alternative population-level evaluation methods. Motivated by the need for robust systems, the TREC Robust Track introduced a number of population-level aggregations capturing how well systems performed on low-performing queries [38]. As part of the TREC Web Track, the risk-sensitive retrieval task evaluated systems according to how they degraded utility relative to a baseline [7]. The retrieval efficiency community includes tail efficiency (e.g., 95th percentile query processing time) in evaluations [27]. Moreover, recent calls from the recommender system community emphasize evaluation of the distribution of measured metric values over a population [13] and from the fairness community to the dis-aggregate metric values across sub-populations [4].

We interrogate average utility and advocate for an alternative population-level evaluation method focused on worst-case analysis, which we refer to as *pessimistic evaluation.* We ground pessimistic evaluation in existing work in equal information access from the information science community, which is based on a larger body of work in fairness. This allows pessimistic evaluation to be based on well-justified methods from political theory. In particular, we introduce the use of lexicographic minimum as a theoretically sound method for pessimistic evaluation.

Our goal is not to demonstrate that pessimistic evaluation dominates existing population-level evaluation based on average system utility. Rather, we will demonstrate that pessimistic evaluation is (i) grounded in ethical and pragmatic concepts, (ii) theoretically complementary to existing robustness and fairness methods, and (iii) empirically validated across a set of retrieval and recommendation tasks. As such, our goal is to demonstrate that pessimistic evaluation complements existing population-level evaluation approaches.

2 Population-level evaluation

Population-level evaluation deals with comparing systems given a set of utility measurements. In this section, we will introduce the formal problem of population-level evaluation and then discuss how we can compare different population-level evaluation methods.

SIGIR-AP '24, December 9–12, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-0724-7/24/12
https://doi.org/10.1145/3673791.3698428

2.1 Preliminaries

For a specific information access task, given an input request x from the space of all requests $\mathcal{X}$, a system f (from the space of all systems $\mathcal{F}$) generates an output (e.g., ranking) with a specific utility (e.g., metric value). An input $x \in \mathcal{X}$ can be a text query, as in traditional information retrieval; a user item history, as in recommendation; or a more complex representation such as, for example, a text query combined with user location information and page view history. For the purpose of our analysis, we only consider measured metric values and are agnostic about whether those utilities are attributed to rankings or strings. So, although a system generates a decision (e.g., a ranking, an answer string), we are only interested in the measured utility, computed by an evaluation metric $\mu : \mathcal{X} \times \mathcal{F} \rightarrow \Re$ defined as a function over the space of all inputs and systems. [1]

In most situations, we have a distribution θ over $\mathcal{X}$ based on users' engagement with the information access system. For example, the probability of a certain input may be proportional to the query frequency in a search engine. In practice, we use a sample of queries $\mathbf{X} \sim \theta$ to evaluate performance. The shorthand $\{\mu\}_{\mathbf{X},f} = \cup_{x \in \mathcal{X}} \mu(x, f)$ refers to the set of measured utilities for a system f over a sample $\mathbf{X}$. We use $n = |\mathbf{X}|$ to refer to the sample size.

2.2 Problem Definition

Determining whether a population-level evaluation is appropriate depends on the population-level objectives of the system designer. In some cases, a quantifiable downstream objective such as revenue can be used to evaluate different population-level evaluation methods. In many cases, though, population-level objectives reflect less well-defined concepts such as fairness or justice. We refer to this class of objectives as population-level normative values of a system.

The goal of population-level evaluation is to sort a set of systems $\mathbf{F} \subseteq \mathcal{F}$ according to the designers population-level objectives. Assume that, for each $f \in \mathbf{F}$, we have measured utility for n inputs. One can approach this problem by defining an *aggregation function* $\overline{\mu} : \Re^n \rightarrow \Re$ that reduces a set of measurements into a single scalar value (e.g., an average). We can then sort $\mathbf{F}$ according to these aggregate scores. Alternative, one can define an *order function* $\overline{\Delta} : \Re^n \times \Re^n \rightarrow \Re$ that generates a scalar value based on a comparison of two sets of n metric measurements. Note that we can derive an order function from an aggregation function but not the other way around.

2.3 Desiderata

Although we can often conduct individual-level metric meta-evaluation by looking at agreement with a ground truth human preference (e.g., 'does the metric ordering agree with a user's preference?'), population-level meta-evaluation leans more heavily on theoretical validation and different methods of empirical validation. In this study, we consider population-level evaluation desiderata based on criteria from measurement theory [18], previously used in the natural language processing [44] and retrieval [11] communities. Similar approaches have been used in ranking metric meta-evaluation [36, 37]. In particular, we will validate a population-level evaluation method using the following criteria.

Content validity. Content validity determines the degree to which the population-level evaluation method is aligned with the theoretical constructs we are interested in measuring. To assess content validity of a method, we determine whether the method is theoretically consistent with normative values of interest. For example, if the system designer is interested in high expected utility for users, then computing the average utility over a set of queries would be consistent with this value; averaging the utility for subscribing users only would ignore the expected impact on users with under-performing queries and not be consistent with the system designer's value.

Convergent validity. Convergent validity determines the degree to which a population-level evaluation method is correlated with other methods supporting the same normative values. To assess the convergent validity of a method by, we measure the empirical correlation between an ordering of $\mathbf{F}$ by the new approach with orderings of $\mathbf{F}$ by existing approaches *for the same normative value.* For example, if both averaging utility over traffic-weighted queries and averaging utility over unique queries aim to capture the expected impact on users, we can measure the correlation between an ordering of $\mathbf{F}$ by averaging traffic-weighted queries with a second ordering by averaging unique queries. Although higher correlation provides evidence that a new measure is consistent with established measures, perfect correlation obviates the need for the new approach.

Discriminant validity. Discriminant validity determines the degree to which a population-level evaluation method is uncorrelated with unrelated methods. To assess the discriminant validity of a method we measure the empirical correlation between an ordering of $\mathbf{F}$ by the new approach with orderings of $\mathbf{F}$ by existing approaches *for different normative values.* In this case, we desire *lower* correlation since it provides evidence that a new measure is different from established measures for different constructs.

Sensitivity. Sensitivity refers to how well a population-level evaluation method can distinguish pairs of systems. A method that is theoretically sound but not sensitive will not be useful in practical settings. We assess the sensitivity of an approach by how often it is unable to distinguish a pair of systems (i.e., the number of tied systems).

3 Properties of Population-Level Evaluation

Since information access systems are tools used by people and a metric measurement reflects the utility of a tool to an individual person, each approach for population-level evaluation makes assumptions about the relative importance of some information needs or people compared to others. As such, we can interpret specific normative values about population-level evaluation as reflecting specific social values. This is consistent with perspectives to information access in information science based on distributive justice

[1] We adopt common practice in search evaluation and assume that systems and metrics are deterministic (i.e., $\mu(x, f)$ will always return the same value). This means that the measured utility will always be the same for the same request.

[28] and echoes recent calls from the recommender system community to reason about the distribution of measured metric values over a population [13] and from the fairness community to the disaggregate metric values across sub-populations [4]. More generally, the recommender systems community is increasingly exploring the incorporation of normative values into design [14, 15, 39–41].

The foundation of a population-level evaluation method is a normative statement about the population-level objective of the designer. While most readers will be familiar with using the arithmetic mean as an aggregation function, in this section, we will introduce three normative statements covering several objectives underlying information access system design. This list is far from exhaustive and the development of appropriate normative values for information access is an active area of research [24].

3.1 Pareto Property

The first property we are interested in is the behavior of a method when a single individual's utility improves. Given two allocations $\{\mu\}_{\mathbf{X},f}$ and $\{\mu\}_{\mathbf{X},f'}$ with equal utility for $|\mathbf{X}| - 1$ individuals, the Pareto property requires that a method prefer the allocation with the higher utility for the remaining individual [20]. Considering the situation where the utility of a single individual improves ensures that an evaluation respects each person's wellbeing. That is, a population-level evaluation that does *not* satisfy this property would sometimes ignore the utility of an individual, even in situations where that of others is not impacted. From the perspective of information access, the Pareto property means that, the performance of all other queries or users being equal, if a system improves the performance for a single query or user, then it should be preferred. The Pareto property is formally defined as,

$$\forall x \in \mathbf{X}, \mu(x,f) \geq \mu(x,f') \land \exists x \in \mathbf{X}, \mu(x,f) > \mu(x,f') \rightarrow f \succ f' \tag{1}$$

For example, while the arithmetic mean of $\{\mu\}_{\mathbf{X},f}$ would satisfy the Pareto property, the median would not, since improving, say, $\min\{\mu\}_{\mathbf{X},f}$ up to the median would not affect the median (or the ordering of systems). While simple, theoretical consistency with the Pareto principle ensures that, no matter the condition, we respect the strict benefit to individuals.

3.2 Average Utilitarianism

While the Pareto property considers a single individual in isolation, we might alternatively consider the expected utility over all individuals. Average utilitarianism prefers the allocation where the expected utility is higher and is well-aligned with empirical risk minimization, the foundation of many machine learning methods. In the context of a commercial information access system, when the measurements are correlated with revenue (or inversely correlated with cost), then average utilitarian is often aligned with cumulative revenue. That said, average utilitarian decision-making focuses on the performance for a random user. In this sense, as a result, if the population is structured so that some inputs or groups of inputs are over-represented, they can dominate the decision-making. Average utilitarianism is formally defined as,

$$\frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mu(x,f) > \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mu(x,f') \rightarrow f \succ f' \tag{2}$$

Since we compare systems over the same population, we can drop the multiplicative factors and recover the utilitarian condition, defined as the cumulative utility over the population,

$$\sum_{x\in\mathcal{X}} \mu(x,f) > \sum_{x\in\mathcal{X}} \mu(x,f') \rightarrow f \succ f' \tag{3}$$

Therefore, we see that average performance has embedded value of utilitarianism [22, 33]. Average utilitarianism is an implicit value present in almost every information access experiment but is justified from a very specific philosophical tradition.

3.3 Difference Principle

While the Pareto property focuses on the difference in utility of an individual when the utility of all other individuals is constant, we often care about the fairness of utility distributed across the population. In particular, Rawls argues that, when an individual does not know which utility in an allocation they will receive, they will rationally decide to prefer the allocation where their worst outcome (i.e., receiving the least utility) is better [31]. This is referred to as the difference principle and underlies many approaches to social justice, including many adopted in the machine learning community [17, 19, 26, 35]. Worst case analysis more generally is useful when evaluating the safety of a system and can provide insight when systems are otherwise difficult to distinguish due to ceiling effects [30].

The difference principle is particularly well-suited for information access evaluation. In the professional librarian community, codes of ethics often include principles of equal access, which are non-utilitarian in nature [21]. Britz [5], based on Rawlsian theories of social justice, advocates for measuring the performance of information access based on those least well-served. In a recommendation context, Singh et al. [34] study worst-case evaluation and optimization to design safe reinforcement learning methods.

The difference principle is formally defined using maximin,

$$\min_{x\in\mathcal{X}}(\mu(x,f)) > \min_{x\in\mathcal{X}}(\mu(x,f')) \rightarrow f \succ f' \tag{4}$$

Unfortunately, because only the minimum value is used, maximin does not satisfy the Pareto principle. Moreoever, the practical problem with this approach is that most systems will have a minimum value of zero, either because systems all fail in different ways or because of outlier contexts that no system can perform well with. Indeed, 83% of runs submitted to the TREC 2021 Deep Learning Passage Ranking task were tied when evaluated using the worst case NDCG at 10. In order to address these issues with maximin, Sen [32] introduced the lexicographic minimum or leximin. Let $\vec{\mu}$ be the measurements $\{\mu\}_{\mathbf{X},f}$ sorted in decreasing order; similarly $\vec{\mu}'$ for $\{\mu\}_{\mathbf{X},f'}$. The leximin preference is defined as,

$$\vec{\mu}_{i^*} > \vec{\mu}'_{i^*} \rightarrow f \succ f' \tag{5}$$

where i^* is the maximum index where $\vec{\mu}_i \neq \vec{\mu}'_i$. Leximin is frequently adopted in the machine learning literature in lieu of maximin [1, 10].

Note that, because it focuses on the single worst utility, the difference principle is not statistical by definition. We contrast this with average utilitarianism, where, because it focuses on the *expected* individual, methods of statistical inference can be leveraged.

3.4 Compatibility between properties

Although these properties are all desirable for different reasons, they are compatible to different degrees. It is easy to confirm that (average) utilitarianism satisfies the Pareto property; all other utilities being equal, increasing the utility to one individual will increase both the total and average utility. On the other hand, by focusing only on the worst case, maximin is indifferent between allocations that would be distinguished by the Pareto principle; if $\{\mu\}_{\mathrm{X},f} = [1, 0.9, 0.1]$ and $\{\mu\}_{\mathrm{X},f'} = [1, 0.8, 0.1]$, then the Pareto property would require that $f \succ f'$ but maximin would be indifferent. Adopting leximin satisfies both the difference principle and the Pareto property. Finally, utilitarianism and the difference principle are incompatible; if $\{\mu\}_{\mathrm{X},f} = [1, 0.0, 0.0]$ and $\{\mu\}_{\mathrm{X},f'} = [0.3, 0.3, 0.3]$, then average utilitarianism would observe $f \succ f'$ but satisfying the difference principles would observe $f \prec f'$

The relationship between these properties clarifies what is valued when we adopt either average utilitarianism or the difference principle. If we adopt average utilitarianism, we cannot provide guarantees about satisfying Rawlsian fairness; if we adopt the difference principle, we cannot provide guarantees about improving average utility. This is important to ensure that evaluation decisions align with normative values and organizational principles. In cases where the information access provider is maximizing engagement, utilitarianism may be more appropriate than the difference principle. In cases where the information access provider seeks to provide equal access, the worst case performance may be more important to consider.

In order to compare the empirical ordering by average utilitarianism and by the difference principle, in Figure 1, we show the rank position of systems evaluated from the arithmetic mean performance to leximin ordering for several datasets. Among a number of small adjustments in rank position, several runs degrade from high positions to very low positions (e.g., 6 → 36, 12 → 32, 16 → 44) and from low positions to very high positions (e.g., 25 → 11, 31 → 9). In general, these results indicate that leximin indeed empirically captures a different phenomenon than arithmetic mean performance.

4 Pessimistic Evaluation

We wish to study population-level evaluation methods consistent with the difference principle, maximin and leximin. That said, given the importance of evaluation in information access problems, several methods exist for measuring the worst-performing queries, often in the context of robustness. In this section, we review these methods.[2] We assume a fixed query sample $\mathbf{X}$ for evaluating systems, consistent with current offline testing practice and, for clarity, write $\{\mu\} = \{\mu\}_{\mathrm{X},f}$ and similarly $\{\mu\}' = \{\mu\}_{\mathrm{X},f'}$.

Geometric mean. Voorhees [38] uses geometric mean utility to emphasize low measured utility. This is formally defined as an aggregation,

$$\mathrm{gavg}(\{\mu\}) = \left(\prod_{\mu \in \{\mu\}} \max(\epsilon, \mu)\right)^{\frac{1}{n}} \tag{6}$$

where ϵ is a small value avoids degenerate effects when $\mu = 0$.[3]

Area under the lower quartile. Voorhees [38] also uses the area under the curve defined by the lowest quartile of queries. This is defined as an aggregation,

$$\mathrm{auc}_4(\{\mu\}) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{j}\sum_{i=1}^{j}\vec{\mu}_{n-i+1} \tag{7}$$

where $k = \lfloor n/4 \rfloor$.

Success at Ten. Voorhees [38] also uses the fraction of queries that have at least one relevant item in the top ten items and is defined as an aggregation,

$$\mathrm{s@10}(\{\mu\}) = \frac{1}{n}\sum_{\mu \in \{\mu\}} \mathrm{I}(\mu > 0) \tag{8}$$

where μ is precision at ten.

Risk-Penalized Gain. Wang et al. [42] uses methods from financial risk modeling to emphasize lower-performing queries. This method compares a baseline allocation to a treatment allocation, measuring average degradation in paired difference in performance. Given two systems f and f', we define the gain of f over f' as an ordering function,

$$\begin{aligned}\mathrm{T}(\{\mu\}, \{\mu\}', \alpha) = &\frac{1}{n}\sum_{x \in \mathbf{X}} \max\left(0, \mu(x, f) - \mu(x, f')\right) \\ &- \frac{1+\alpha}{n}\sum_{x \in \mathbf{X}} \max\left(0, \mu(x, f') - \mu(x, f)\right)\end{aligned} \tag{9}$$

where $\alpha \geq 0$ and we recover ordering by the arithmetic mean when $\alpha = 0$. From a preference perspective, the gain is asymmetric, meaning that $\mathrm{T}(\{\mu\}, \{\mu\}', \alpha) \neq -\mathrm{T}(\{\mu\}', \{\mu\}, \alpha)$. To address this, we combine the two directions as $\overline{\mathrm{T}}(\{\mu\}, \{\mu\}', \alpha) = \mathrm{T}(\{\mu\}, \{\mu\}', \alpha) - \mathrm{T}(\{\mu\}', \{\mu\}, \alpha)$.

Gini Coefficient. The Gini coefficient is a measure of the average difference between all pairs of utilities in an allocation. Although not used for population-level evaluation in the ranking literature,[4] the Gini coefficient is often used in the economics literature to quantify inequity in distribution of utility and is defined as an aggregation,

$$\mathrm{gini}(\{\mu\}) = \frac{1}{2n^2}\sum_{\mu, \mu' \in \{\mu\}} \frac{|\mu - \mu'|}{\overline{\{\mu\}}} \tag{10}$$

where $\overline{\{\mu\}}$ is the arithmetic mean of $\{\mu\}$. The Gini coefficient is bounded between 0 and 1, with 0 reflecting maximum equality and 1 reflecting maximum inequality. Note that the Gini coefficient *only* measures inequality and does not capture utility of an allocation.

5 Theoretical Analysis

Understanding whether a particular population-level evaluation method is consistent with a property in Section 3 is important for several reasons. First, because information access systems are increasingly subject to regulation, grounding evaluation in clear

[2] We omit the plethora of fair ranking metrics because they focus on fair exposure of providers rather than utility to users.

[3] Voorhees [38] sets this value to 0.00001, which we adopt in our experiments.

[4] The Gini coefficient *has* been used to measure the inequality of retrieval of items in both search [3] and recommendation [12, 25].

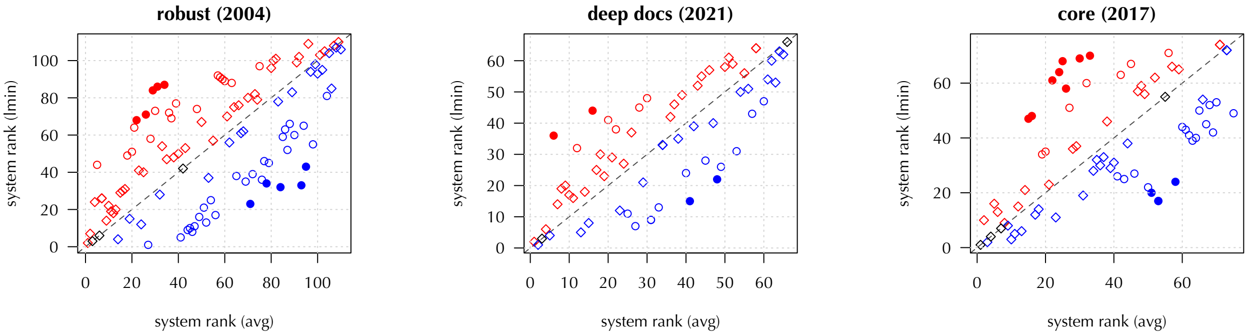


Figure 1: Ranking of runs in three TREC tracks (see Table 4) according to associated metrics when ordering systems by average performance (horizontal) and leximin (vertical). ⋄: change in position less than one quintile. ○: change in position between one and two quintiles. ●: change in position greater than two quintiles. Red: degradation in ranking. Blue: improvement in rank position. Black: no change in rank position.

conceptual terms provides transparency and theoretical guarantees about the alignment with philosophical and legal principles (e.g., fairness) [9]. The value of transparency and these theoretical guarantees extends beyond external regulatory agencies to internal stakeholders, including engineering and policy teams within organization [8]. Second, theoretically characterization extends axiomatic analysis of individual-level evaluation to population-level evaluation [2, 6, 29].

As noted in Section 3.4, although both maximin and leximin satisfy the difference principle by construction, only leximin satisfies the Pareto principle. In this section, we will examine whether the methods in Section 4 are consistent with the properties in Section 3.

Geometric mean. The geometric mean is equivalent to applying a logarithmic transformation to measurements before computing the arithmetic mean. Because the logarithm is monotonically increasing in utility, if $\{\mu\}$ and $\{\mu\}'$ only differ in a single individual who has higher utility in $\{\mu\}'$, then the arithmetic mean of the logarithms will be greater for $\{\mu\}'$. Next, although equivalent to the arithmetic mean of the logarithms, this does not ensure that the geometric mean is consistent with average utilitarianism. Consider $\{\mu\} = [1, 0.9, 0.1]$ and $\{\mu\}' = [0.5, 0.5, 0.5]$ where the geometric mean prefers $\{\mu\}'$ while the arithmetic mean prefers $\{\mu\}$. Moreover, despite emphasizing lower performance, the geometric mean is not consistent with the difference principle. If $\{\mu\} = [0.25, 0.25, 0.25]$ and $\{\mu\}' = [1, 0.9, 0.1]$, the geometric mean prefers $\{\mu\}'$ while the difference principle would select $\{\mu\}$.

Area under the lower quartile. Because the area under the lower quartile only considers a subset of utility values, any change in utility occurs above the lower quartile is ignored. This means that it does not satisfy the Pareto property. This also means that it is not consistent with average utilitarianism. Finally, because the area under the curve accumulates averages within the bottom quartile, situations can arise which are inconsistent with the difference principle. For example, consider $\{\mu\} = [1, 0.9, 0.7, 0.6, 0.4, 0.3, 0.1, 0.05]$ and $\{\mu\}' = [1, 0.9, 0.7, 0.6, 0.4, 0.3, 0.3, 0.0]$, where the area under the lower quartile prefers $\{\mu\}'$ while the difference principle would select $\{\mu\}$.

Success at Ten. While perhaps correlated with many ranking measures, success at ten does not provide any guarantee about utility when considering other measures that inspect beyond the tenth position. If utility is defined by query-level success at ten, then the population-level success at ten reduces to the arithmetic mean, satisfying the Pareto property and, trivially, average utilitarianism. Because query-level success at ten is binary, if population-level success at ten detects a difference between two systems, it will be equivalent to the preference detected by leximin, satisfying the difference principle. That said, for other metrics, population-level success at ten provides no guarantees.

Risk-Penalized Gain. The symmetric version of gain, $\overline{\mathrm{T}}$, reduces to the difference in arithmetic mean performance with a multiplicative factor of $2 + \alpha$, indicating that it satisfies the Pareto property and average utilitarianism but not the difference principle.

Gini Coefficient. Because the Gini coefficient focuses on measuring equality, it prefers allocations that are more uniform, resulting in inconsistency with the Pareto property. For example, if $\{\mu\} = [0.6, 0.5, 0.5]$ and $\{\mu\}' = [0.5, 0.5, 0.5]$, the Gini coefficient would prefer $\{\mu\}'$ while the Pareto property would lead to $\{\mu\}$ being preferred. The same example can be used to demonstrate that the Gini coefficient is not consistent with average utilitarianism. We can see that Gini coefficient is also inconsistent with the difference principle by considering $\{\mu\} = [0.8, 0.6, 0.5, 0.3]$ and $\{\mu\}' = [0.5, 0.3, 0.3, 0.2]$, where the Gini coefficient would prefer $\{\mu\}'$ while the difference principle would prefer $\{\mu\}$.

Summary. We summarize results in Table 1. These analyses all compromise the content validity of these population-level evaluation methods from the perspective of the properties discussed in Section 3. This does not suggest that the methods are not useful in evaluating systems, but that they are inconsistent with various properties important to designing information access systems. In Section 6, we will move from theoretical analyses to study the empirical behavior of these methods.

Table 1: Summary of population-level evaluation properties: Pareto property (PP), average utilitarianism (AU), difference principle (DP) for minimum (min), leximin (lmin), arithmetic mean (avg), geometric mean (gavg), success at ten (s@10), area under the lower quartile (auc$_4$), risk-penalized gain (gain).

	PP	AU	DP
min	✗	✗	✓
lmin	✓	✗	✓
avg	✓	✓	✗
gavg	✓	✗	✗
s@10	✗	✗	✗
auc$_4$	✗	✗	✗
gain	✓	✓	✗
gini	✗	✗	✗

6 Empirical Analysis

Having examined the properties of several population-level evaluation methods, we now turn to studying the empirical relationship between these methods and leximin. Our objective in this section is to measure the convergent validity, discriminant validity, and sensitivity of leximin. In order to measure convergent validity, we compute Kendall's τ between the ranking of systems ordered by leximin and the ranking of systems ordered by other pessimistic evaluation methods (i.e., those in Section 4). We adopt τ_b to handle ties [23]. In order to measure discriminant validity, we compute Kendall's τ between the ranking of systems ordered by leximin and the ranking of systems ordered by other methods *not* focused on pessimistic evaluation. In this case, we consider the arithmetic mean and leximax, which is analogous to leximin but starts at the best-case utility. In order to measure sensitivity, we count the number of tied systems under each population-level evaluation method.

6.1 Data

We analyzed population-level evaluation across a wide range of information access tasks where using the difference principle is well-motivated (Section 3). For information retrieval contexts, we used official runs submitted to 24 different TREC tracks. For recommendation contexts, we used publicly available runs for three recommendation tasks: movielens, beerAdvocate, and libraryThing [37]. Dataset details are available in Appendix A.

We focused analysis on comparing the following population-level methods: minimum, leximin, arithmetic mean, geometric mean, success at ten, and lowest quartile area under the curve. We omit symmetric risk-penalized gain because it is equivalent to the arithmetic mean and the Gini coefficient because it does not capture total utility.

6.2 Results

Results for corpora using average precision as the utility measure are presented in Table 2. Results for other metrics on the robust (2004) corpus are presented in Table 3.

When measuring convergent validity, we are interested in higher correlation with population-level methods intended to capture the same higher level concept. In Table 2, when comparing leximin to minimum, geometric mean, success at ten, and lowest quartile area under the curve, we observe τ values in general ranging from 0.50 to 0.90, suggesting high empirical agreement. Cases where correlations are weaker occur when the population-level evaluation method incurs substantial ties (e.g., web (2009), deep-docs (2020), legal (2007)). Although leximin and minimum both capture the difference principle, we observe a correlation lower than 1 because leximin computes preferences even when we observe a tie in minimum utilities. The next highest correlation after minimum is with the geometric mean, which is the only other pessimistic evaluation method that satisfies the Pareto property. This means that, unlike success at ten and lowest quartile area under the curve, like leximin, it considers the full set of queries.

When measuring discriminant validity, we are interested in lower correlation with population-level methods intended to capture the different higher level concepts, in our case average utilitarianism and best-case utility. When comparing leximin to arithmetic mean, we observe τ values in general in the range from 0.20 to 0.60, suggesting low to moderate correlation. Comparing the τ across methods for a single condition (i.e., row), we see that the correlation with the arithmetic mean is lower than the correlations with pessimistic methods. Moreover, the correlation is consistently weakest with leximax, which captures the best-case performance and should be low.

Table 3 demonstrates that our observations about convergent and discriminant validity are consistent across other metrics. We noticed that, especially for metrics with rank cutoffs (e.g., ndcg@10, p@10), the success at ten method correlated much higher with leximin.

Finally, when inspecting the number of ties for each method, leximin always is tied for the lowest value and is comparable to other methods with high sensitivity. This is especially salient when comparing leximin with minimum, the only other method consistent with the difference principle, which consistently demonstrates tied performance. As such, if we are interested in satisfying the difference principle and conducting experiments, leximin should be preferred to minimum.

7 Discussion

We were motivated to rigorously define and understand pessimistic evaluation from basic concepts grounded in normative principles of fairness and information access. Our theoretical results demonstrate that leximin is the only population-level evaluation in our suite that satisfies the fundamental Pareto property and difference principle. Our empirical results demonstrate the convergent and discriminant validity of leximin while also having high sensitivity.

We based our adoption of the difference principle on work from information science advocating for Rawlsian fairness in information access [5, 21] and separate work in the machine learning community [17, 19, 26, 35]. As such, providing formal guarantees of population-level methods aligning with it is important, especially in responsible artificial intelligence contexts where stakeholders from multiple disciplines need to align [8].

Table 2: Kendall's τ for various datasets using average precision as the utility metric. Comparison between system ranking based on leximin (lmin) and minimum score (min), geometric mean (gavg), success at ten (s@10), area under the lower quartile (auc$_4$), arithmetic mean (avg), and leximax (lmax). Number of ties in parentheses. Dashes reflect situations where all systems are tied according to the method. Bold: highest τ in convergent validity analysis. Italics: lowest correlation in discriminant validity analysis.

		pessimistic evaluation						
	nruns	lmin	min	gavg	s@10	auc$_4$	avg	lmax
robust (2004)	110	1.000 (2)	**0.882** (54)	0.665 (2)	0.566 (95)	0.682 (2)	0.474 (2)	*0.193* (2)
core (2017)	75	1.000 (6)	**0.998** (16)	0.557 (6)	0.481 (73)	0.592 (6)	0.433 (6)	*0.331* (6)
core (2018)	72	1.000 (2)	**0.987** (16)	0.629 (2)	0.603 (67)	0.664 (9)	0.512 (2)	*0.450* (2)
web (2009)	48	1.000 (0)	0.287 (46)	0.761 (0)	0.666 (32)	**0.785** (5)	0.475 (0)	*0.043* (0)
web (2010)	32	1.000 (2)	**0.970** (12)	0.556 (2)	0.523 (28)	0.661 (2)	0.244 (2)	*0.071* (2)
web (2011)	61	1.000 (8)	**0.863** (39)	0.733 (8)	*0.372* (57)	0.671 (8)	0.640 (8)	0.444 (8)
web (2012)	48	1.000 (4)	**0.957** (22)	0.586 (4)	0.321 (42)	0.648 (4)	0.401 (4)	*0.302* (4)
web (2013)	61	1.000 (8)	**0.974** (30)	0.564 (8)	0.506 (57)	0.708 (8)	0.370 (8)	*0.151* (8)
web (2014)	30	1.000 (4)	**0.992** (10)	0.473 (4)	0.508 (26)	0.621 (4)	0.386 (4)	*0.095* (4)
deep-docs (2019)	38	1.000 (0)	**0.967** (10)	0.366 (0)	0.527 (36)	0.616 (0)	0.260 (0)	*-0.084* (0)
deep-docs (2020)	64	1.000 (0)	0.539 (54)	**0.619** (0)	0.545 (62)	0.563 (0)	0.408 (0)	*0.277* (0)
deep-docs (2021)	66	1.000 (6)	**0.975** (21)	0.520 (6)	0.641 (65)	0.604 (6)	0.210 (6)	*-0.041* (6)
deep-docs (2022)	42	1.000 (0)	**0.883** (20)	0.847 (0)	0.697 (37)	0.844 (0)	0.784 (0)	*0.526* (0)
deep-docs (2023)	5	1.000 (0)	**0.949** (2)	0.600 (0)	0.527 (2)	0.600 (0)	0.600 (0)	*0.400* (0)
deep-pass (2019)	37	1.000 (0)	0.549 (31)	**0.628** (0)	0.563 (35)	0.532 (0)	0.580 (0)	*0.517* (0)
deep-pass (2020)	59	1.000 (0)	**0.808** (35)	0.615 (0)	0.523 (49)	0.617 (3)	0.501 (0)	*0.416* (0)
deep-pass (2021)	63	1.000 (0)	0.589 (51)	**0.704** (0)	0.519 (57)	0.642 (2)	0.520 (0)	*0.051* (0)
deep-pass (2022)	100	1.000 (0)	**0.793** (66)	0.731 (0)	0.632 (92)	0.722 (0)	0.649 (0)	*0.457* (0)
deep-pass (2023)	35	1.000 (0)	**0.824** (22)	0.771 (0)	0.755 (27)	0.768 (0)	0.681 (0)	*0.382* (0)
legal (2006)	34	1.000 (0)	- (34)	**0.722** (0)	0.433 (25)	0.712 (6)	0.490 (0)	*0.005* (0)
legal (2007)	68	1.000 (0)	**0.943** (25)	0.514 (0)	*0.323* (60)	0.433 (3)	0.430 (0)	0.347 (0)
podcasts (2020)	14	1.000 (0)	0.711 (10)	**0.912** (0)	0.769 (7)	0.739 (4)	0.868 (0)	*0.560* (0)
podcasts (2021)	27	1.000 (11)	0.649 (21)	0.725 (11)	**0.770** (21)	0.704 (12)	0.642 (11)	*0.487* (11)
tot (2023)	30	1.000 (2)	- (30)	**0.714** (2)	0.414 (8)	0.503 (26)	0.341 (2)	*0.304* (2)
movielens	21	1.000 (0)	- (21)	0.752 (0)	0.790 (0)	**0.802** (4)	0.676 (0)	*0.600* (0)
beerAdvocate	21	1.000 (0)	- (21)	**0.857** (0)	0.781 (0)	- (21)	0.771 (0)	*0.705* (0)
libraryThing	21	1.000 (0)	- (21)	**0.952** (0)	0.933 (0)	0.900 (8)	0.914 (0)	*0.838* (0)

Table 3: Kendall's τ for robust (2004) using R-Precision (rp), average precision (ap), normalized discounted cumulative gain (ndcg), precision (p), and reciprocal rank (rr). Formatting identical to Table 2.

	pessimistic evaluation						
	lmin	min	gavg	s@10	auc$_4$	avg	lmax
rp	1.000 (2)	- (110)	0.840 (2)	0.778 (95)	**0.848** (4)	0.516 (2)	*0.192* (2)
ap	1.000 (2)	**0.882** (54)	0.665 (2)	0.566 (95)	0.682 (2)	0.474 (2)	*0.193* (2)
ndcg	1.000 (2)	**0.882** (58)	0.687 (2)	0.526 (95)	0.723 (2)	0.524 (2)	*0.178* (2)
ndcg@100	1.000 (2)	0.352 (103)	**0.802** (2)	0.688 (95)	0.787 (2)	0.494 (2)	*0.233* (2)
p@100	1.000 (2)	0.347 (110)	0.788 (2)	0.688 (95)	**0.853** (4)	0.471 (2)	*0.005* (2)
ndcg@10	1.000 (2)	- (110)	0.890 (2)	**0.987** (95)	0.908 (4)	0.530 (2)	*0.141* (2)
p@10	1.000 (2)	- (110)	0.909 (2)	**0.987** (95)	0.962 (41)	0.526 (2)	*0.174* (2)
rr	1.000 (2)	**0.882** (64)	0.703 (2)	0.589 (95)	0.660 (2)	0.527 (2)	*0.472* (2)

Although we have grounded pessimistic evaluation in arguments from information science, alternative positions may demand alternative population-level evaluation methods. Just as average utilitarianism can be justified for commercial organizations, other properties can be justified by other contexts. For example, in some commercial situations, there is a privileged group of subscribed customers that deserves substantially more attention during population-level evaluation to ensure retention [43]. Or, alternative notions of justice may require entirely new population-level fairness evaluation formalisms [16].

As discussed in Section 3.3, leximin is not statistical by definition since it emphasizes the absolute worst case instead of the expected case. This does not imply that there is not uncertainty in estimating worst-case performance differences since it can arise from query sampling, randomness in system decisions, in addition to other sources.

One way to introduce uncertainty is to define a relaxed version of leximin. For example, we can use a moving average to smooth measurements. For a lag of k, we compare $\sum_{j=0}^{k-1} \vec{\mu}_{i+j}$ and $\sum_{j=0}^{k-1} \vec{\mu}'_{i+j}$ instead of $\vec{\mu}_i$ and $\vec{\mu}'_i$. This means i iterates from $n-(k-1)$ to 1, recovering the arithmetic mean when $k=n$. In Figure 2, we demonstrate how adjusting k generates rankings of systems smoothly transitioning between the difference principle (low values of k) and average utilitarianism (high values of k). This behavior is similar to lower quartile area under the curve except that, like the leximin, it backs off to higher quantiles in the presence of a tie. Smoothed leximin satisfies the Pareto property because, if $\{\mu\}$ and $\{\mu\}'$ only differ in a single individual who has higher utility in $\{\mu\}'$, then each $\sum_{j=0}^{k-1} \vec{\mu}_{i+j} \leq \sum_{j=0}^{k-1} \vec{\mu}'_{i+j}$ and at least one inequality is strict. Moreoever, it is easy to see that, unless $k=n$, smoothed leximin evaluation can be inconsistent with the average utilitarian decision. As an example, consider $\{\mu\} = [1, 0.0, 0.0]$ and $\{\mu\}' = [0.2, 0.2, 0.2]$ where, unless $k=n$, smoothed leximin prefers $\{\mu\}'$. Similarly, unless $k=1$, smoothed leximin evaluation can be inconsistent with the difference principle. As an example, consider $\{\mu\} = [0.1, 0.1, 0.1]$ and $\{\mu\}' = [0.25, 0.25, 0.0]$ where, unless $k=1$, smoothed leximin prefers $\{\mu\}'$. We believe smoothed leximin provides one way to begin to explore statistical methods for pessimistic evaluation.

While leximin provides a lens into worst-case performance, there are situations when aggregation-based methods are more appropriate. For example, we may be interested in a single score assigned to each system for a downstream process or decision. Aggregation-based methods may also obviate the need for leximin in some metric conditions. For example, leximin becomes increasingly similar to the arithmetic mean as the number of discrete utility values decreases. When there are two values—as with success at ten—leximin and the arithmetic mean are equivalent. To see how discretization of utility values affects the relationship between leximin and the arithmetic mean, we conducted the following experiment. Under the first discretization method, we progressively set all utility valued below a threshold to 0. Under the second discretization method, we removed significant digits from the utility value. Figure 3 shows the Kendall's τ between the leximin ordering and the arithmetic mean ordering as we discretized average precision values. Under both discretization methods, we observe a gradual convergence with the arithmetic mean. This suggests that, even if we are interested in the difference principle, there are situations where the arithmetic mean is equivalent.

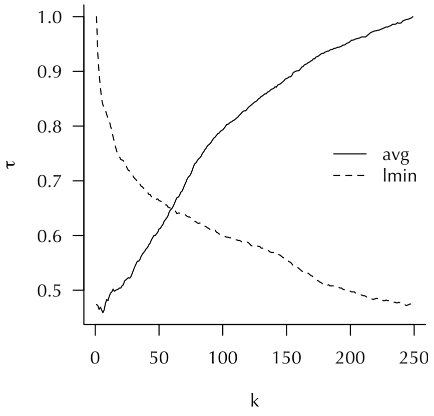
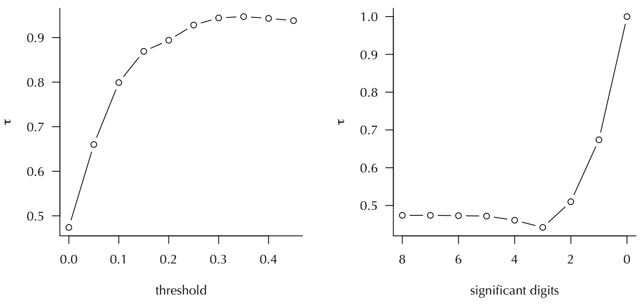


Figure 2: Kendall's τ between system orderings by smoothed leximin and arithmetic mean (solid) and leximin (dashed) using average precision on Robust 2004 (see Section 6.1 for details).



(a) **(b)**

Figure 3: Discretization of average precision values for Robust 2004. (a) Metric values below the threshold set to zero. (b) Metric values quantized to a specific number of significant digits.

At a higher level, we hope that a rigorous conceptual and theoretical grounding of population-level evaluation provides a template for others interested in developing new methods. In particular, because population-level evaluation often explicitly or implicitly captures social values, understanding those foundations when analyzing existing or designing new methods is important for policy, legal, and ethical integrity.

8 Conclusion

While the information retrieval community has a rich history of individual-level evaluation research focused around metric design, substantially less work exists studying population-level evaluation, even though it is fundamental to all retrieval experiments. As an example of the importance of analysis of population-level evaluation,

Table 4: Datasets used in empirical analysis. Runs submitted to the associated TREC track or recommendation task.

	requests	runs	rel/request	docs/request
robust (2004)	249	110	69.93	913.82
core (2017)	50	75	180.04	8853.11
core (2018)	50	72	78.96	7102.61
web (2009)	50	48	129.98	925.31
web (2010)	48	32	187.63	7013.21
web (2011)	50	61	167.56	8325.07
web (2012)	50	48	187.36	6719.53
web (2013)	50	61	182.42	7174.38
web (2014)	50	30	212.58	6313.98
deep-docs (2019)	43	38	153.42	623.77
deep-docs (2020)	45	64	39.27	99.55
deep-docs (2021)	57	66	189.63	98.83
deep-docs (2022)	76	42	1245.62	100
deep-docs (2023)	82	5	75.10	100
deep-pass (2019)	43	37	95.40	892.51
deep-pass (2020)	54	59	66.78	978.01
deep-pass (2021)	53	63	191.96	99.95
deep-pass (2022)	76	100	1315.22	100
deep-pass (2023)	82	35	103.18	100
legal (2006)	39	34	110.85	4835.07
legal (2007)	43	68	101.02	22240.30
podcasts (2020)	48	14	43.67	963.40
podcasts (2021)	50	27	30.80	781.15
tot (2023)	150	31	1	1000
movielens	6005	21	18.87	100.00
libraryThing	7227	21	13.15	100.00
beerAdvocate	17564	21	13.66	99.39

we introduced pessimistic evaluation through leximin, a method firmly grounded in robust philosophical and moral traditions. We contrast the guarantees provided by this theoretical foundation with related methods from robustness measures in information retrieval. We further demonstrated content, convergent, and discriminant validity of leximin as well as a competitive sensitivity. We advocate information retrieval experimenters—especially those in organizers with values aligned with the difference principle—to complement existing population-level methods with leximin.

A Data

All TREC runs and relevance judgments were downloaded from NIST.[5] Recommendation runs and judgments were downloaded from a public repository.[6] Datasets are detailed in Table 4. Metrics were computed using the official trec_eval package.[7]

References

[1] Jacob Abernethy, Robert E. Schapire, and Umar Syed. 2024. Lexicographic Optimization: Algorithms and Stability. *Proceedings of Machine Learning Research* 238 (2024), 2503–2511.

[2] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *The 41st International ACM SIGIR Conference on Research &#38; Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 625–634.

[3] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 561–570.

[4] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 368–378. https://doi.org/10.1145/3461702.3462610

[5] Johannes J. Britz. 2004. To Know or not to Know: A Moral Reflection on Information Poverty. *Journal of Information Science* 30, 3 (2004), 192–204.

[6] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval (ICTIR '13)*. Association for Computing Machinery, New York, NY, USA, 22–29.

[7] Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles L. A. Clarke, and Ellen M. Vorhees. 2014. TREC 2013 Web Track Overview. In *Proceedings of the 22nd Text REtrieval Conference (TREC 2013)* (proceedings of the 22nd text retrieval conference (trec 2013) ed.). https://www.microsoft.com/en-us/research/publication/trec-2013-web-track-overview/

[8] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 227–236.

[9] Tommaso Di Noia, Nava Tintarev, Panagiota Fatourou, and Markus Schedl. 2022. Recommender systems under European AI regulations. *Commun. ACM* 65, 4 (mar 2022), 69–73.

[10] Emily Diana, Wesley Gill, Ira Globus-Harris, Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Lexicographically Fair Learning: Algorithms and Generalization. In *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference (LIPIcs, Vol. 192)*, Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 6:1–6:23.

[11] Fernando Diaz and Bhaskar Mitra. 2024. Recall, Robustness, and Lexicographic Evaluation. arXiv:2302.11370 [cs.IR]

[12] Virginie Do and Nicolas Usunier. 2022. Optimizing Generalized Gini Indices for Fairness in Rankings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 737–747.

[13] Michael D. Ekstrand, Ben Carterette, and Fernando Diaz. 2024. Distributionally-Informed Recommender System Evaluation. *ACM Trans. Recomm. Syst.* 2, 1, Article 6 (mar 2024), 27 pages. https://doi.org/10.1145/3613455

[14] Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2022. Measuring Commonality in Recommendation of Cultural Content: Recommender Systems to Enhance Cultural Citizenship. In *Proceedings of the 16th ACM Conference on Recommender Systems.*

[15] Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2024. Measuring Commonality in Recommendation of Cultural Content to Strengthen Cultural Citizenship. *ACM Trans. Recomm. Syst.* 2, 1 (mar 2024).

[16] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (2022), 90.

[17] Yash Gupta, Runtian Zhai, Arun Suggala, and Pradeep Ravikumar. 2023. Responsible AI (RAI) Games and Ensembles. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 72717–72749.

[18] David J. Hand. 2010. *Measurement Theory and Practice: The World Through Quantification.* Wiley.

[19] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 1929–1938.

[20] Iwao Hirose. 2014. The Structure of Aggregation. In *Moral Aggregation.* Oxford University Press.

[21] Anna Lauren Hoffmann. 2017. Beyond distributions and primary goods: Assessing applications of rawls in information science and technology literature since 1990. *Journal of the Association for Information Science and Technology* 68, 7 (2017), 1601–1618.

[22] T. M. Hurka. 1982. Average Utilitarianisms. *Analysis* 42, 2 (1982), 65–69. http://www.jstor.org/stable/3327924

[5] https://trec.nist.gov/results.html

[6] https://github.com/dvalcarce/evalMetrics

[7] https://github.com/usnistgov/trec_eval

[23] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.

[24] Johannes Kruse, Lien Michiels, Alain Starke, Nava Tintarev, and Sanne Vrijenhoek. 2024. NORMalize: A Tutorial on the Normative Design and Evaluation of Information Access Systems. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 422–424.

[25] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszár, and Rumman Chowdhury. 2022. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns* 3, 8 (2024/07/06 2022).

[26] Mike Li, Hongseok Namkoong, and Shangzhou Xia. 2021. Evaluating model performance under worst-case subpopulations. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 17325–17334.

[27] Joel Mackenzie and Alistair Moffat. 2020. Examining the Additivity of Top-k Query Processing Innovations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1085–1094.

[28] Kay Mathiesen. 2015. Informational Justice: A Conceptual Framework for Social Justice in Library and Information Services. *Library Trends* 64, 2 (2015).

[29] Javier Parapar and Filip Radlinski. 2021. *Towards Unified Metrics for Accuracy and Diversity for Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 75–84.

[30] Christian W. Probst, Andreas Gal, and Michael Franz. 2005. Average case vs. worst case: margins of safety in system design. In *Proceedings of the 2005 Workshop on New Security Paradigms (NSPW '05)*. Association for Computing Machinery, New York, NY, USA, 25–32.

[31] John Rawls. 1971. *A Theory of Justice: Original Edition*. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9z6v

[32] Amartya Sen. 1970. *Collective Choice and Social Welfare*. Holden-Day.

[33] Henry Sidgwick. 2011. *The Methods of Ethics*. Cambridge University Press.

[34] Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, Ed H. Chi, Jilin Chen, and Alex Beutel. 2020. Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach. In *3rd FAccTRec Workshop: Responsible Recommendation*.

[35] Nikolaj Thams, Michael Oberst, and David Sontag. 2022. Evaluating Robustness to Dataset Shift via Parametric Robustness Sets. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 16877–16889.

[36] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the Robustness and Discriminative Power of Information Retrieval Metrics for Top-N Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 260–268.

[37] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal* 23, 4 (2020), 411–448.

[38] E.M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*.

[39] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 208–219.

[40] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 173–183.

[41] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Jordi Viader Guerrero, Alain Starke, and Nava Tintarev (Eds.). 2023. *Proceedings of the First Workshop on Normative Design and Evaluation of Recommender Systems*.

[42] Lidan Wang, Paul N. Bennett, and Kevyn Collins-Thompson. 2012. Robust Ranking Models via Risk-Sensitive Optimization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 761–770. https://doi.org/10.1145/2348283.2348385

[43] Chuhan Wu, Qinglin Jia, Zhenhua Dong, and Ruiming Tang. 2023. Customer Lifetime Value Prediction: Towards the Paradigm Shift of Recommender System Objectives. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1293–1294.

[44] Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10967–10982.