An Outcome Test of Discrimination for Ranked Lists

Jonathan Roth
jonathanroth@brown.edu
Brown University
Providence, RI, USA

Guillaume Saint-Jacques
guillaume.saintjacques@gmail.com
Apple
USA

YinYin Yu
yinyyu@linkedin.com
LinkedIn
USA

ABSTRACT

This paper extends Becker [3]'s outcome test of discrimination to settings where a (human or algorithmic) decision-maker produces a ranked list of candidates. Ranked lists are particularly relevant in the context of online platforms that produce search results or feeds, and also arise when human decisionmakers express ordinal preferences over a list of candidates. We show that non-discrimination implies a system of moment inequalities, which intuitively impose that one cannot permute the position of a lower-ranked candidate from one group with a higher-ranked candidate from a second group and systematically improve the objective. Moreover, we show that that these moment inequalities are the *only* testable implications of non-discrimination when the auditor observes only outcomes and group membership by rank. We show how to statistically test the implied inequalities, and validate our approach in an application using data from LinkedIn.[1]

ACM Reference Format:
Jonathan Roth, Guillaume Saint-Jacques, and YinYin Yu. 2022. An Outcome Test of Discrimination for Ranked Lists. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3531146.3533102

1 INTRODUCTION

Researchers are often interested in testing whether a human or algorithmic decision-maker is biased against members of a protected group (e.g. race or gender). Substantial attention has been paid to the case where the decision is binary – e.g., whether or not to grant a loan, accept a student to college, hire a job candidate, etc. However, in a variety of relevant domains, the decision-maker produces a *ranked list* of candidates. Ranked lists are particularly relevant in the context of online platforms: LinkedIn provides recruiters with an ordered list of candidates, Google returns an ordered list of search results, and Facebook and Twitter provide users with an ordered feed of posts. Ranked lists are relevant in other domains, as well: for example, hospitals participating in the National Residency Match Program (NRMP) provide a listwise ranking of candidate residents [19], and experiments in behavioral economics have asked participants to rank which of the other participants they would like to be grouped with [6].

We consider the setting where an Auditor observes ranked lists of candidates produced by a Ranker. The Auditor observes each candidate's group status G and outcome Y. Importantly, the Auditor does not observe all of the features X that are available to the decision-maker. This reflects a realistic limitation to (human or algorithmic) audits in a variety of domains. When the decision-maker is human, it is nearly always the case that there are factors that are observed by the decision-maker but not by the auditor – for example, an auditor of the NRMP would not be able to observe everything that occurred during the candidate's interview. Likewise, external auditors of tech platforms will almost never have access to all of the features used by the algorithm. Even internally within tech platforms it is often difficult to retrospectively reconstruct all of the features used by an algorithm, since all of the relevant user data may not be saved for privacy reasons. Moreover, even if all the data were observed, the covariates may be so high-dimensional that it is difficult to condition on the full set of covariates in any practical analysis.

We then ask how the Auditor can test whether the Ranker is biased against a protected group in forming their rankings. Our notion of bias extends Becker [3]'s notion of taste-based discrimination to the context of listwise rankings. In particular, we will say that the Ranker is unbiased if they sort candidates to maximize an objective function that values placing candidates with better outcomes earlier in the list. This notion of unbiasedness is referred to as "accurate statistical discrimination" in the economics literature. In the computer science literature, a Ranker would be said to be unbiased if they order the candidates based on the predictions of a Bayes-optimal score. As we show, this form of objective nests optimizing the Net Discounted Cumulative Gain (NDCG) objective commonly used for search algorithms. It can also be motivated by a simple model in which the objective is to maximize total engagement, and engagement with a post is an increasing function of its quality and rank in the list.

Our first main theoretical result is that the null hypothesis of no bias implies a system of conditional moment inequalities. Intuitively, these moments impose that whenever we see a particular configuration of the candidates (e.g. a woman first, man second, etc.), we should not be able to flip the order of some of the candidates and improve the objective function on average. For instance, we should not be able to increase the objective on average by flipping the position of the first two candidates whenever the first-ranked candidate is male and the second is female.

Our second main theoretical result is that this system of moment inequalities is a *sharp* testable implication of the hypothesis of no bias. Specifically, whenever the moment inequalities are satisfied,

[1] LinkedIn is engaged in a long-term effort to measure the impact of our products and ensure that they do not reinforce existing social inequalities. To learn more about these ongoing efforts, please visit: https://lnkd.in/equitable-outcomes-meaning. Prior research showed that the InstaJobs algorithm can function to lower inequality broadly between different groups of members [20].

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533102

there exists a distribution for the unobserved covariates such that the observed data corresponds with the utility maximization of an unbiased decision maker.

Our theoretical results allow us to leverage a large econometrics literature on testing moment inequalities to develop statistical tests of bias in settings with list-wise rankings (see Canay and Shaikh [5], Molinari [15] for reviews of the moment inequality literature). We discuss several practical considerations, including: reducing the dimension of the large number of implied inequalities; adjusting for position effects – wherein a candidate's realized position has a causal effect on their outcome; and incorporating observed features about the candidates.

Finally, we showcase our proposed procedure in a validation exercise using data from the InstaJobs algorithm at LinkedIn. The InstaJobs algorithm is an algorithm for determining whether to send users a notification about a job they may be interested in. The algorithm generates a predicted score for each candidate, and sends notifications to candidates above a threshold. We use the scores constructed by the algorithm for each job to create a listwise ranking of the candidates, and apply our proposed tests to this ranking data. This construction allows us to validate the listwise outcome test by comparing its results to what we would obtain by directly examining the score used to generate the ranking (which is not typically available for a ranking algorithm in practice). We find using the listwise outcome test that when a female and male candidate are in adjacent ranks, the *lower-ranked* male candidate systematically has better outcomes than the *higher-ranked* female candidates. This finding accords with a direct examination of the scores used for rankings, which show that the algorithm is under-calibrated for men relative to women. To be clear, these results suggest that the algorithm is not sorting candidates based on the average outcome (a weighted average of job applications and recruiter responses) considered in this paper — however, this is potentially consistent with the algorithm maximizing other objectives or satisfying other notions of fairness (e.g. accounting for the fact that recruiters themselves may be less likely to respond to woman than men with the same "objective" qualifications).

Our results relate to a large literature on detecting discrimination in economics, computer science, and other fields. See Lang and Spitzer [14] for a recent review of the economics literature on discrimination, which has primarily focused on the case where the decision-maker makes individual-level decisions (e.g. give out a loan, hire a candidate) rather than produces listwise rankings.[2] Several notions of fairness for listwise ranking algorithms have been considered previously in the computer science literature, as reviewed in Pitoura et al. [16]. Multiple papers have considered demographic parity constraints, which require that exposure (i.e. the distribution of rankings) be similar across groups [7, 9, 10, 21, 22]. Other work has considered the notions of disparate treatment and disparate impact, which restrict that exposure be proportional to average group-level utility or outcomes [21]. Beutel et al. [4] propose a notion of fairness that extends the notion of equal opportunity [11] to the listwise ranking setting: this requires that the probability that a candidate is ranked below another candidate with a worse outcome does not differ across groups. The notion of fairness we consider here is distinct, and is based on the question of whether the ranking is consistent with maximizing an objective function that does not depend on group status directly. That is, is the Ranker ranking candidates using a Bayes optimal score given the Ranker's information set? As discussed in Corbett-Davies and Goel [8] and Rambachan et al. [17] in the context of binary classification problems, decision rules that maximize an unbiased utility function may violate demographic parity or equalized odds if the distribution of risks differs across populations, a problem known as *inframarginality*. Similar distinctions arise in the context of listwise rankings.

2 MODEL

The model we consider consists of a Ranker and an Auditor. The Ranker – which could be either an algorithm or human – observes an unordered list of candidates and their characteristics, and produces a ranked list of the candidates. The Auditor then observes the ranked list of candidates and their outcomes, and wants to test whether the Ranker is biased.

2.1 Set-up

Data-generating Process. The Ranker is presented with queries indexed by q in which they are asked to rank J candidates with characteristics $X_{1q}, ..., X_{Jq}$ and group status $G_{1q}, ..., G_{Jq}$. We denote by $\mathcal{I}_q = \{(X_{jq}, G_{jq}\}_{j=1}^{J}$ the information provided to the Ranker for query q. After observing $\mathcal{I}_q$, the Ranker produces a ranked list of the candidates – formally this is a map $j_q : \{1, ..., J\} \to \{1, ..., J\}$ where $j_q(r)$ corresponds with the index of the candidate in rank r. Rank 1 is the best rank, and we suppose that there are no ties in the rankings, so that j_q is one-to-one. After the candidates are ranked, they realize outcomes $Y_{1q}, ..., Y_{Jq}$. We suppose for now that the outcomes Y_{jq} do not depend on the rankings — in Section 4.1 below, we show that the framework can accommodate certain forms of "position effects", wherein the observed outcome is affected by the ranking itself; in this case, Y_{jq} corresponds with the "position-adjusted" outcome for unit j in query q.

Auditor. There is an Auditor tasked with evaluating whether the Ranker is biased. For each query q, the Auditor sees the rank-ordered list of outcomes $Y_q = (Y_{j_q(1)q}, ..., Y_{j_q(J)q})$ as well as the group statuses by rank, $G_q = (G_{j_q(1)q}, ..., G_{j_q(J)q})$. Importantly, the Auditor does not see the characteristics $X_q = (X_{j_q(1)q}, ..., X_{j_q(J)q})$ used to form the rankings. We show that similar results arise if the Auditor observes a subset of the variables in X_q in Section 4.1.

2.2 Tests of Unbiasedness

We now describe the notion of unbiasedness that we will test, which extends the logic of Becker [3] to the setting of list-wise rankings. Specifically, we will be interested in testing the hypothesis that the Ranker chooses the ranking $j_q(\cdot)$ to maximize

$$E\left[\sum_r w_r Y_{j_q(r)q} \mid \mathcal{I}_q\right], \tag{1}$$

where the w_r are a strictly decreasing sequence of positive weights.

[2] One exception to this is Castillo and Petrie [6], who conduct a lab experiment in which participants rank whom they would like to be grouped with. Castillo and Petrie [6] focus on differences in average ranks across groups.

Definition 1. We say that the Ranker is unbiased if they choose $j_q(\cdot)$ to maximize (1) for a decreasing sequence of positive weights w_r.

Intuitively, the fact that the w_r are decreasing means that the Ranker prefers to place candidates with higher values of Y earlier in the list. This corresponds with expected utility maximization if the Ranker's utility function is $U(Y_q, G_q) = \sum_r w_r Y_{j_q(r)q}$, which depends only on the rank-ordered outcomes for the candidates, and not directly on their group-status G_q. That is, an unbiased ranker performs what is called "accurate statistical discrimination" in the economics literature. We note that the researcher need not specify the w_r; our test will be valid for the null hypothesis that the Ranker maximizes (1) for *any* decreasing sequence of w_r.

It is straightforward to show that an unbiased Ranker chooses the order $j_q(\cdot)$ that corresponds with sorting the candidates based on their expected outcomes given the Ranker's information set ($E[Y_{jq}|\mathcal{I}_q]$). That is, an unbiased Ranker forms the *Bayes optimal score* $E[Y_{jq}|\mathcal{I}_q]$, and then orders the candidates based on their score. This is sometimes referred to as the Probability Ranking Principle [18, 21].

Example 1 (NDCG). A common objective function used for ranking algorithms is Net Discounted Cumulative Gain, abbreviated NDCG [12]. Intuitively, NDCG is a weighted average of outcomes by position, normalized by the score that would be obtained if all candidates were sorted perfectly. Formally, NDCG is defined as the ratio $DCG/IDCG$, where

$$DCG = \sum_r Y^*_{j_q(r)}/log_2(r+1),$$

Y^*_j is a relevance score, and $IDCG$ is the idealized value of DCG that would be realized if the candidates had been sorted perfectly in decreasing order of Y^*,

$$IDCG = \sum_r Y^*_{j^*(r)}/log_2(r+1),$$

where $j^*(r)$ is the index with the rth largest value of Y^*_j. It is then apparent that maximizing expected NDCG is equivalent to maximizing an objective of the form (1), with $Y_j = Y^*_j/IDCG$, and $w_r = log_2(r+1)^{-1}$.

▲

Example 2 (Maximizing total engagement). It is well-known that users tend to engage more with posts earlier in a list than later in the list. Suppose that placing a post one position later in the list causally reduces the amount of user engagement by the factor $1+\gamma$. (The parameter γ could be estimated in an experiment that randomizes the order of the candidates.) Then for $w_r = 1/(1+\gamma)^r$, we have that the total number of clicks generated by a given ranking is $\sum_r w_r Y_{j_q(r)q}$, where Y_{jq} is the number of clicks that candidate j would have received if placed in the first position in the query (and thus is not affected by rank).[3] Thus, the objective function above corresponds with maximizing total engagement after accounting for the causal effect of position on engagement. ▲

2.2.1 *Possible violations of the null.*

We now discuss several possible deviations from the null hypothesis of unbiasedness. One important type of violation of this hypothesis is when the Ranker maximizes expected utility for the utility function

$$U(Y_q, G_q) = \sum_r w_r (Y_{j_q(r)q} - \tau G_{j_q(r)}), \tag{2}$$

so that the outcome is effectively penalized by τ for candidates from group $G = 1$ relative to group $G = 0$. When the decision-maker is human, such a utility function may arise owing to racial animus against the $G = 1$ group, which Becker [3] referred to as taste-based discrimination. With an algorithm, such bias might arise if the algorithm's score is "boosted" to try to increase exposure for particular groups of candidates.

Another important possible type of violation of the null is if an algorithm (or human perception) is trained on selected training data. In this case, the algorithm may be approximately maximizing $\tilde{E}\left[\sum_r w_r Y_{j_q(r)q} \mid \mathcal{I}_q\right]$, where $\tilde{E}[\cdot]$ denotes the expectation in the training sample. If the training sample is very different from the target population, then the algorithm that sorts based on the Bayes score using $\tilde{E}[\cdot]$ will generally not equal the optimizer of (1), so that the null will be violated.

A third important case where the null will be violated is if there is *omitted payoffs bias* [13]. That is, the Ranker may optimize an objective like (1) for some $Y^* \neq Y$. For example, in the context of hiring, Y^* may be actual productivity on the job, whereas Y may be some proxy such as the score on a performance review. It is thus important to realize that our null hypothesis is that the Ranker maximizes the expected utility based on the chosen outcome Y. A violation of the null could correspond with biased predictions on the part of the Ranker, *or* with accurate maximization of a different objective. This is particularly important for interpretation of our test in settings where racial or gender bias may affect the measured label Y; if the Ranker is instead maximizing an alternative outcome Y^* (e.g. a "debiased" outcome), then this would lead to a violation of the null. For example, if Y corresponds with whether a recruiter on a job-search platform clicks on a particular candidate's profile, then the null of our test is that the algorithm sorts candidates by their probability of being clicked. If, in fact, recruiters have animus against candidates from particular demographic groups — that is, they are less likely to click on candidates from certain groups conditional on "objective" features on their resume — then our notion of unbiasedness may not correspond with the Ranker sorting candidates based only on their objective qualifications for the job. The choice of Y is thus an important element of the test.

Finally, we note that the alternative hypothesis of our tests encompasses many other possible deviations as well. For example, the null that the Ranker maximizes (1) could also be violated if the Ranker sorts by some estimated score that is noisier from some groups than it is for others. In our view, one advantage of our test is that the alternative hypothesis encompasses many interesting violations of "accurate statistical discrimination." In many cases where the null is violated, it may be interesting to understand *which* type of violation of the null occurred, which we think is an interesting topic for future research.

[3] If in practice candidate j receives engagement Y^*_{jq} and is placed in position a, we can construct $Y_{jq} = (1+\gamma)^a Y^*_{jq}$ as the "position adjusted outcome."

3 THEORETICAL RESULTS

We now provide two main theoretical results. First, we show that unbiasedness by the Ranker implies a system of conditional moment inequalities. Second, we show that these moment inequalities are a sharp implication of unbiased behavior.

Proposition 3.1. *If the Ranker is unbiased, then for all $a < b$,*

$$E[Y_{j_q(a)q} - Y_{j_q(b)q} \,|\, G_q = g] \geq 0 \tag{3}$$

for all g such that $P(G_q = g) > 0$.

Proof. Consider the counterfactual assignment rule that permutes $j_q(a)$ and $j_q(b)$ whenever $G_q = g$, and otherwise corresponds with the observed choice rule. Denote the rankings of this rule by $\tilde{j}_q(\cdot)$. Note that if $j_q(\cdot)$ maximizes $E\left[\sum_r w_r Y_{j_q(r)q} \,|\, \mathcal{I}_q\right]$ for all $\mathcal{I}_q$, then by iterated expectations it must maximize $E\left[\sum_r w_r Y_{j_q(r)q}\right]$. Then, the difference in objective value from using the observed choice rule versus the permuted choice rule is given by

$$\begin{aligned} &E\left[\textstyle\sum_r w_r Y_{j_q(r)q}\right] - E\left[\textstyle\sum_r w_r Y_{\tilde{j}_q(r)q}\right] \\ &= P(G_q = g)(w_a - w_b) E[Y_{j_q(a)q} - Y_{j_q(b)q} \,|\, G_q = g]. \end{aligned}$$

If the observed choice rule is optimal, then the expression in the previous display must be non-negative. However, $P(G_q = g)(w_a - w_b) > 0$ by assumption, from which the result follows. □

It is straightforward to show that the change in the objective from permuting the candidates in positions a and b is proportional to $Y_{j_q(a)q} - Y_{j_q(b)q}$. Proposition 3.1 thus intuitively states that if the Ranker is unbiased, then we shouldn't be able to permute the position of the candidates in positions a and b whenever we see $G_q = g$ and improve the objective. In other words, higher ranked candidates should always have higher values of Y on average, regardless of the group orientation of the query.

Example 3. Suppose that $J = 2$ and in every query there is one man and one woman, so that $G = (1, 0)$ or $G = (0, 1)$. Then (3) is equivalent to

$$E[Y_{j_q(1)q} - Y_{j_q(2)q} \,|\, G_1 = 1, G_2 = 0] \geq 0$$
$$E[Y_{j_q(1)q} - Y_{j_q(2)q} \,|\, G_1 = 0, G_2 = 1] \geq 0$$

The first inequality says that the outcome for the higher-ranked candidate should be larger on average when the higher-ranked candidate is male (and hence the lower-ranked candidate is female), whereas the second inequality is analogous for the case where the higher-ranked candidate is female. ▲

Remark 1 (Sufficiency of adjacent ranks)**.** Proposition 3 is stated in terms of comparisons of all ranks (a, b) with $a < b$. It suffices to consider adjacent ranks, i.e. pairs of the form $(a, b) = (k, k + 1)$. This is because if the inequality in (3) holds for $(a, b) = (k, k + 1)$ and $(a, b) = (k + 1, k + 2)$, then adding the two inequalities implies that it also holds for $(a, b) = (k, k + 2)$, and so on.

Our next result formalizes the notion that the inequalities in Proposition 3.1 are the only testable implication of unbiasedness by the Ranker. It states that if the observed data satisfies the moment inequalities in Proposition 3.1, then there exists a latent distribution for the covariates X_q such that the observed distribution corresponds with unbiased behavior by the decision-maker.

Proposition 3.2. *Suppose the inequalities in Proposition 3.1 are satisfied. Then there exists a distribution for $\mathcal{I}_q$ such that the observed distribution (Y_q, G_q) corresponds with the decision rule that maximizes $E\left[\sum_r w_r Y_{j_q(r)q} \,|\, \mathcal{I}_q\right]$.*

Proof. We construct a distribution for $\mathcal{I}_q$ that satisfies the proposition. Intuitively, we construct $\mathcal{I}_q$ such that whenever the decisionmaker chooses $G_q = g$, their expectation for the candidate in position r is precisely $E[Y_{j_q(r)q} \,|\, G_q = g]$, and thus the observed ranking is optimal since this expectation is monotonically decreasing in the rank. Formally, let $\mathcal{I}_q^G = \{G_{jq}\}_{j=1}^J$ and $\mathcal{I}_q^X = \{X_{jq}\}_{j=1}^J$. Construct $\mathcal{I}_q^G$ to have the distribution corresponding with $\{G_q\}$, where $\{G_q\}$ denotes the unordered list of elements in G_q. Let $\mathcal{I}_q^X$ have support that is one-to-one with the support of G_q. In a slight abuse of notation, we will write $\mathcal{I}_q^X = g$ to denote that $\mathcal{I}_q^X$ takes the value in its support mapping to g. We construct $\mathcal{I}_q^X$ such that $P(\mathcal{I}_q^X = g \,|\, \mathcal{I}_q^G = \{g\}) = P(G_q = g \,|\, \{G_q\} = \{g\})$. Next, we construct $Y_q \,|\, \mathcal{I}_q^X = g$ to have the same distribution as $Y_q \,|\, G_g = g$. Equation (3) then implies that if $r_1 < r_2$, then $E[Y_{j_q(r1)q} - Y_{j_q(r2)q} \,|\, \mathcal{I}_q^X = g] \geq 0$, and thus j_q maximizes the objective. Moreover, by construction of the conditional probabilities, the implied distribution of (Y_q, G_q) matches that in the data. □

Remark 2 (Relationship to marginal outcome tests)**.** Unbiased behavior by the ranker implies that $E[Y_{j_q(a)} | \mathcal{I}_q]$ should be *equal* to $E[Y_{j_q(a+1)} | \mathcal{I}_q]$ if the Ranker is indifferent between the candidates in positions a and $a + 1$ (marginal candidates). However, since it is not observed which candidates are marginal, the only testable implications involve comparisons between adjacent ranks, even when these are *inframarginal*, meaning that the ranker strictly prefers the candidate in rank a to the one in $a + 1$. As a result, the inequality in (3) may be strict if the Ranker is unbiased – that is, candidates ranked in position a may be strictly better on average than candidates ranked in position $a + 1$. By continuity arguments, it follows that if the Ranker has a small amount of bias (e.g. $\tau > 0$ is very small in (2)), then the inequalities in equation (3) might still be satisfied, so that discriminatory behavior is not detectable. In this case, the Ranker would not be maximizing (1), but the testable implications of unbiasedness given in Proposition 3.1 would still hold – i.e. the test would have no power to detect the violation of the null of unbiasedness. However, this inframarginality problem should be less severe when many candidates are observed, since the expected difference between the a and $(a + 1)$-th best candidates should be small. For example, if $E[Y_{jq} | I_q]$ is $i.i.d.$ uniformly distributed across j, then for an unbiased ranker $E[Y_{j_q(a)q} - Y_{j_q(a+1)q} | I_q]$ is $O_P(1/J)$ uniformly in a.[4]

4 TESTING

We now discuss how one can test the hypothesis that a Ranker is unbiased given a sample of queries $q = 1, ..., Q$. For simplicity, we will focus on the case where the queries q are i.i.d., although the approach we describe will extend easily to clustered or weakly dependent data.

[4] This follows from the fact that the difference in consecutive order statistics of the uniform distribution is distributed $Beta(1, N)$.

Pointwise tests. We first note that for fixed values of a, b and g, the hypothesis that equation (3) holds is simply the hypothesis that the population mean of $Y_{j_q(a)q} - Y_{j_q(b)q}$ is larger than zero among the population of queries with $G_q = g$. This individual hypothesis can be tested with a standard one-sided t-test for the mean of the population with $G_q = g$. Such tests will be asymptotically valid under standard regularity conditions that allow for an application of a central limit theorem. These individual tests for fixed values of (a, b, g) are useful in that they can help identify where (if anywhere) the Ranker appears to be making biased decisions, which may be useful in addressing any detected bias.

Joint tests. Although the individual tests described above will be valid for each individual hypothesis for a fixed (a, b, g) combination, it is well known that there is a problem of multiple hypothesis testing if such tests are conducted for each possible value of (a, b, g). Fortunately, a large literature in econometrics has developed joint tests for a system of moment inequalities such as (1) for all relevant (a, b, g); see, for instance, Canay and Shaikh [5], Molinari [15] for recent reviews.

4.1 Implementation and Extensions

Choice of Moments. One important practical point for implementation is that the dimensionality of the vector G_q can be quite large in practice: for example, if g takes on two values and there are 30 candidates in the query, then there are $2^{30} \approx 10^9$ possible values of G_q. To reduce the dimensionality in our implementation below, when comparing the outcomes of rank a to rank b, we condition on the group membership of the candidates in ranks a and b (i.e. $G_{q,a}, G_{q,b}$), but not on the group status of the other candidates in the list, which substantially reduces the dimensionality. Although in theory this may reduce the power of the test, we suspect that most of the pertinent information about the outcomes for ranks a and b is captured by their own group status, and not by the group status of other candidates in the query. In our implementation we also only test moments comparing adjacent ranks, which as discussed in Remark 1, is equivalent to the null hypothesis across all ranks.

Position effects. Our analysis so far has assumed that the outcome for candidate j does not depend on their rank in the query. In practice, however, there may be causal effects of position on the outcome, e.g. the same candidate may get more clicks if ranked higher in a search on LinkedIn. If the Auditor knows the causal effect of position – e.g., from an experiment that randomizes search order – the inequalities can be adjusted to account for this. In particular, suppose the Auditor knows that putting a candidate in position a increases outcomes by a factor of $(1 + \gamma)$ relative to ranking the same candidate in position b. Then the change in the objective from swapping the candidate in positions a and b would be proportional to $(1 + \gamma)Y_{j_q(b)q} - Y_{j_q(a)q}$, i.e. a comparison of the outcomes that each candidate would have reached if they had been placed in position b. Optimizing behavior by the Ranker implies that this swap can't improve the objective, and thus yields the inequality

$$E[Y_{j_q(a)q} - (1 + \gamma)Y_{j_q(b)q} \mid G_q = g] \geq 0 \tag{4}$$

instead of (3). Note that this is equivalent to testing (3) where the outcome used is the "position-adjusted" outcome rather than the observed outcome, as in Example 2.[5] We note that if the relevance score Y is positive and $\gamma \geq 0$ (so earlier positions are better), then equation (4) implies (3). Thus, tests of (3), which ignore position effects, will still be valid for the null of unbiasedness, but they may have lower power to detect violations of the null if position effects are important.

Partially observed features. Our analysis so far has assumed that the features X are completely unobserved by the auditor. In practice, a subset of the features used by the decisions-maker may be observed: that is, $X_q = (X_q^O, X_q^U)$, where X_q^O are features observed by the auditor and X_q^U are unobserved features. In this case, the same arguments above can be applied within each group of candidates with the same observable features – i.e, conditional on X_q^O. Thus, the sharp testable implication of unbiased rankings in this case is that

$$E[Y_{j_q(a)q} - Y_{j_q(b)q} \mid G_q = g, X_q^O = x] \geq 0$$

for almost-every (g, x). Such inequalities can be tested using methods for *conditional* moment inequalities [1].

Mitigation. Our focus is on detecting bias in rankings, and we primarily leave the problem of mitigating bias to future work. However, we note that examining which moments appear to be violated (i.e., which permutations of ranks could improve the objective), may help algorithm designers to identify what features of the algorithm deserve additional scrutiny.

5 VALIDATION USING LINKEDIN DATA

We now provide a validation exercise using data from LinkedIn's InstaJobs algorithm.[6]

5.1 InstaJobs

Background. InstaJobs is an algorithm that sends LinkedIn members (candidates) a notification about a job posting that they may be interested in.[7] The algorithm uses features about the job posting and the candidates to predict the probability the candidate will apply for the job as well as the probability the application will receive attention from the recruiter. Specifically, the algorithm scores candidates based on the predicted value for the outcome

$$\begin{aligned} Y^*_{jq} &= \alpha * 1[\textit{job applied}]_{jq} + 1[\textit{job applied} \\ &\quad \& \textit{ application received recruiter attention}]_{jq} \end{aligned} \tag{5}$$

where $\alpha \in (0, 1)$ is a (proprietary) scalar parameter that determines the relative weight placed on applications versus recruiter attention.

[5] For simplicity, our discussion above assumes that there are constant proportional position effects, so that every candidate's outcome in position a is $(1 + \gamma)$ times higher than their outcome if they'd been placed in position b. This constant effects assumption can be relaxed. The key restriction is that $-E[Y_{j_q(a)q} - (1 + \gamma)Y_{j_q(b)q} \mid G_q = g]$ corresponds with the change in objective from flipping candidates a and b when $G_q = g$. A sufficient condition is that percentage change in the average outcome from moving a candidate from position b to a does not depend on G_q or j_q; we can thus accommodate heterogeneous proportional position effects so long as they are not systematically correlated with j_q or G_q.

[6] In part based on the results in this paper, the version of this algorithm studied in this paper has subsequently been deprecated.

[7] Here "candidates" refers to LinkedIn members who are candidates for receiving a notification regarding a job posting.

All candidates with a score above a certain threshold are sent a notification.

Creating Listwise Rankings. Importantly, InstaJobs is not a listwise ranking algorithm, but rather a pointwise classification algorithm that creates a score and makes a binary decision based on this score. However, we can generate data *as if* InstaJobs were a list-wise ranking algorithm by rank-ordering the candidates for each job by the score the InstaJob algorithm assigned them. An advantage of this approach is that, since we have access to the true scores underlying the InstaJobs algorithm, we can validate the listwise approach by investigating whether it produces similar answers to a direct investigation of the scores generating the rankings. Also, since InstaJobs is not actually a ranking algorithm, its results should not be affected by position bias.

Practical Implementation. We have data on the InstaJobs algorithm scores for approximately 193,000 jobs. We create a ranked list of the top 11 candidates for each job based on the InstaJobs score.[8] (We use the top 11 candidates so that we can make 10 comparisons between adjacent ranks.) We then implement the listwise outcome test, using gender as the group variable and NDCG as the objective (see Example 1). We construct tests for each individual comparison using one-sided t-tests, and joint tests based on moment inequality tests with least-favorable critical values that assume all moments have mean-zero (e.g. Andrews and Soares [2]).[9] For comparison, we also can look directly at whether the score is calibrated to check whether marginally-notified candidates have the same outcomes regardless of their gender, where the outcome is the algorithm's objective function given in equation (5).

Results. Figure 1 shows estimates of the moments comparing adjacent ranks. The left panel compares the average value of Y for women in one rank relative to men in the adjacent rank below, with the x-axis denoting the rank of the woman. We see that the point estimate is negative for all ranks, indicating that the lower-ranked men actually have *better* outcomes than the women ranked immediately above them, which means that we could improve the objective function (NDCG) by swapping their ranks. Moreover, the differences are individually statistically significant for 8 of the 10 ranks, and we can jointly reject the hypothesis that all of differences are non-negative ($p < 0.01$; see Table 1 for exact p-values). The right panel of Figure 1 makes similar comparisons, except between higher-ranked men and lower-ranked women, and finds only positive differences, indicating that higher-ranked men do indeed have higher values of Y than women ranked below them. The listwise outcome test thus suggests that the algorithm is systematically ranking men below women despite having better outcomes. Since our data allows us to look at the scores generating the ranks, we can validate the listwise outcome test by looking at the outcomes by score directly: the binned scatter plot in Figure 2 shows that men with a given score do indeed have systematically higher outcomes than women with the same score, confirming the conclusion of the listwise outcome test.[10]

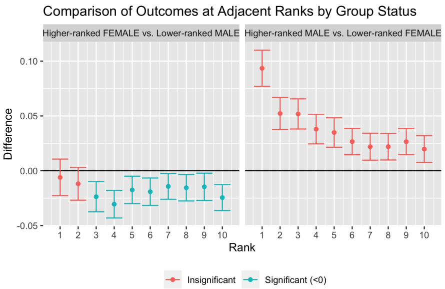


Figure 1: Moments Comparing Men and Women in Adjacent Ranks
Note: This figure shows comparisons of the average outcome Y between candidates in a given rank relative to the candidate in the next rank. The left panel shows these differences for queries where the higher-ranked candidate is female and the lower-ranked candidate is male. The right panel shows analogous comparisons when the higher-ranked candidate is male and the lower-ranked candidate is female. The x-axis denotes the rank of the higher-ranked candidate in the comparisons (where 1 is the highest rank).

6 CONCLUSION

This paper considers how to test for bias when the decision-maker (which could be a human or algorithm) produces a listwise ranking. Our notion of unbiasedness corresponds with "accurate statistical discrimination", i.e. ranking by a Bayes optimal score. We show that a sharp testable implication of unbiased behavior is a system of moment inequalities, and discuss how these can be tested in practice. We validate the methodology using the InstaJobs algorithm

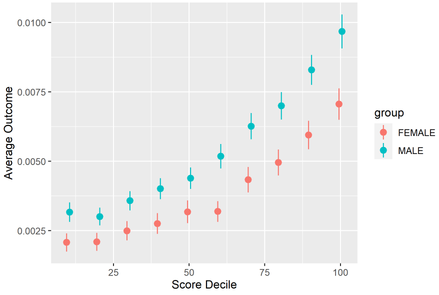
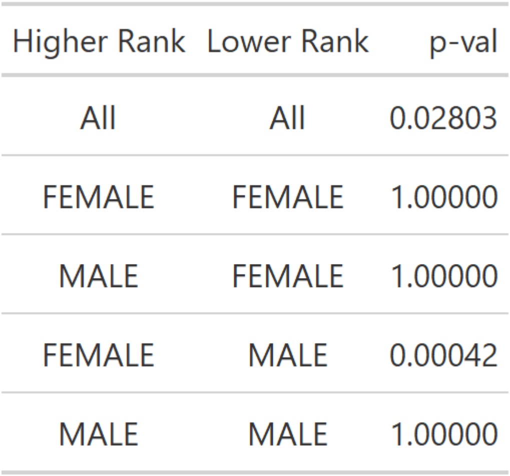


Figure 2: Pointwise Comparison of Outcome by Score
Note: this figure shows a binned scatterplot of the outcome used by the InstaJobs algorithm against the InstaJobs algorithm score. The series are separated for male and female candidates.

[8] We restrict only to candidates who receive a notification, since outcomes are only available for these candidates. Thus, some queries will have fewer candidates if fewer than 11 people were notified for a given job.

[9] Formally, we create unconditional moments of the form $E\left[\left(Y_{j_q(a)q} - Y_{j_q(a+1)q}\right) 1[G_{q,a} = g_1, G_{q,a+1} = g_2]\right] \geq 0$, where Y is the relevance score Y^* normalized by the IDCG, as in Example 1.

[10] This is true even if we impose a linear control for variation of scores within each score decile.



Higher Rank	Lower Rank	p-val
All	All	0.02803
FEMALE	FEMALE	1.00000
MALE	FEMALE	1.00000
FEMALE	MALE	0.00042
MALE	MALE	1.00000

Table 1: p-values for Joint Hypotheses
Note: This table shows p-values for the joint hypothesis that all of the moments comparing higher-ranked to lower-ranked candidates are positive. The top row uses all of the moments, whereas the second only considers moments comparing women in adjacent ranks, the next row compares higher-ranked men to lower-ranked women, and so on. The p-values are constructed using least-favorable critical values for moment inequalities.

at LinkedIn. In future work, we plan to apply this methodology to other listwise rankings algorithms at LinkedIn.

REFERENCES

[1] Donald W. K. Andrews and Xiaoxia Shi. 2013. Inference Based on Conditional Moment Inequalities. *Econometrica* 81, 2 (2013), 609–666. https://doi.org/10.3982/ECTA9370 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9370.

[2] Donald W. K. Andrews and Gustavo Soares. 2010. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica* 78, 1 (2010), 119–157. https://doi.org/10.3982/ECTA7502

[3] Gary Becker. 1957. *The Economics of Discrimination.* University of Chicago Press.

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

[5] Ivan A Canay and Azeem M Shaikh. 2017. Practical and theoretical advances in inference for partially identified models. *Advances in Economics and Econometrics* 2 (2017), 271–306.

[6] Marco Castillo and Ragan Petrie. 2010. Discrimination in the lab: Does information trump appearance? *Games and Economic Behavior* 68, 1 (Jan. 2010), 50–59. https://doi.org/10.1016/j.geb.2009.04.015

[7] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2017. Ranking with Fairness Constraints. (April 2017). https://arxiv.org/abs/1704.06840v4

[8] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (2018), 25.

[9] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (July 2019), 2221–2231. https://doi.org/10.1145/3292500.3330691 arXiv: 1905.01989.

[10] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. (May 2021). https://doi.org/10.1145/3404835.3462850

[11] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]* (Oct. 2016). http://arxiv.org/abs/1610.02413 arXiv: 1610.02413 version: 1.

[12] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/345508.345545

[13] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (Feb. 2018), 237–293. https://doi.org/10.1093/qje/qjx032

[14] Kevin Lang and Ariella Kahn-Lang Spitzer. 2020. Race Discrimination: An Economic Perspective. *Journal of Economic Perspectives* 34, 2 (May 2020), 68–89. https://doi.org/10.1257/jep.34.2.68

[15] Francesca Molinari. 2020. Chapter 5 - Microeconometrics with partial identification. In *Handbook of Econometrics*, Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin (Eds.). Handbook of Econometrics, Volume 7A, Vol. 7. Elsevier, 355–486. https://doi.org/10.1016/bs.hoe.2020.05.002

[16] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *arXiv:2104.05994 [cs]* (April 2021). http://arxiv.org/abs/2104.05994 arXiv: 2104.05994.

[17] Ashesh Rambachan, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. An Economic Perspective on Algorithmic Fairness. *AEA Papers and Proceedings* 110 (May 2020), 91–95. https://doi.org/10.1257/pandp.20201036

[18] S.E. ROBERTSON. 1977. THE PROBABILITY RANKING PRINCIPLE IN IR. *Journal of Documentation* 33, 4 (Jan. 1977), 294–304. https://doi.org/10.1108/eb026647 Publisher: MCB UP Ltd.

[19] Alvin E. Roth. 1984. The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *Journal of Political Economy* 92, 6 (Dec. 1984), 991–1016. https://doi.org/10.1086/261272 Publisher: The University of Chicago Press.

[20] Guillaume Saint-Jacques, Amir Sepehri, Nicole Li, and Igor Perisic. 2020. Fairness through Experimentation: Inequality in A/B testing as an approach to responsible design. *arXiv:2002.05819 [cs, econ]* (Feb. 2020). http://arxiv.org/abs/2002.05819 arXiv: 2002.05819.

[21] 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*. ACM, London United Kingdom, 2219–2228. https://doi.org/10.1145/3219819.3220088

[22] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. (June 2017). https://doi.org/10.1145/3132847.3132938