On the Impact of Outlier Bias on User Clicks

Fatemeh Sarvi
AIRLab, University of Amsterdam
Amsterdam, The Netherlands
f.sarvi@uva.nl

Ali Vardasbi
University of Amsterdam
Amsterdam, The Netherlands
a.vardasbi@uva.nl

Mohammad Aliannejadi
University of Amsterdam, Amsterdam
The Netherlands
m.aliannejadi@uva.nl

Sebastian Schelter
University of Amsterdam, Amsterdam
The Netherlands
s.schelter@uva.nl

Maarten de Rijke
University of Amsterdam, Amsterdam
The Netherlands
m.derijke@uva.nl

ABSTRACT

User interaction data is an important source of supervision in counterfactual learning to rank (CLTR). Such data suffers from presentation bias. Much work in unbiased learning to rank (ULTR) focuses on position bias, i.e., items at higher ranks are more likely to be examined and clicked. Inter-item dependencies also influence examination probabilities, with *outlier* items in a ranking as an important example. Outliers are defined as items that observably deviate from the rest and therefore stand out in the ranking. In this paper, we identify and introduce the bias brought about by outlier items: users tend to click more on outlier items and their close neighbors.

To this end, we first conduct a controlled experiment to study the effect of outliers on user clicks. Next, to examine whether the findings from our controlled experiment generalize to naturalistic situations, we explore real-world click logs from an e-commerce platform. We show that, in both scenarios, users tend to click significantly more on outlier items than on non-outlier items in the same rankings. We show that this tendency holds for all positions, i.e., for any specific position, an item receives more interactions when presented as an outlier as opposed to a non-outlier item. We conclude from our analysis that the effect of outliers on clicks is a type of bias that should be addressed in ULTR. We therefore propose an outlier-aware click model that accounts for both outlier and position bias, called outlier-aware position-based model (OPBM). We estimate click propensities based on OPBM; through extensive experiments performed on both real-world e-commerce data and semi-synthetic data, we verify the effectiveness of our outlier-aware click model. Our results show the superiority of OPBM against baselines in terms of ranking performance and true relevance estimation.

CCS CONCEPTS

• **Information systems** → *Learning to rank*.

KEYWORDS

Click bias; Outliers; Counterfactual learning to rank

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591745

ACM Reference Format:
Fatemeh Sarvi, Ali Vardasbi, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2023. On the Impact of Outlier Bias on User Clicks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591745

1 INTRODUCTION

Ranking systems optimize ranking decisions to increase user satisfaction. Implicit user feedback is an important source of supervision that reflects the preferences of actual users. However, user interaction data (e.g., clicks) suffers from presentation bias, which can make its naïve use as training data highly misleading [18].

Much work in unbiased learning to rank (ULTR) focuses on position bias [3, 17, 19, 43], i.e., the phenomenon that higher-ranked results are more likely to be examined and thus clicked by users [17] than lower-ranked results. Besides position there are several other factors that affect users' examination model and clicks [1, 11, 27, 34, 44]. Previous work has shown that inter-item dependencies can influence user judgments of relevance and the examination order of items [12, 34, 44]. The existence of *outlier* items is a specific case of inter-item dependencies [34]. Sarvi et al. [34] define outliers in a ranking as items that observably deviate from the rest of the list w.r.t. item features, such that they stand out and catch users' attention. For instance, in an e-commerce search scenario, if only one item on the page features a "Best Seller" tag, it can be considered as an outlier, because the tag differentiates it from the rest of the items in the ranking, thereby attracting users' attention.

Outlier bias. An outlier in a list of items can alter the examination probabilities, such that the probability of examination is higher for the outlier item (if it exists) and its neighboring items than the probability assigned by the position bias assumption [34].

Although it has been shown that outliers affect examination probabilities [34], their impact on user click behavior is unknown. In this work, we hypothesize that clicks are biased by the existence of outliers. We refer to this phenomenon as *outlier bias* and aim to understand and address this effect. To begin, we conduct a user study where we compare the click-through rate (CTR) for specific items in two conditions: once shown as outliers and once as non-outlier items in the list. We find that users behave differently in relation to an item given its outlierness condition. The CTR of a specific item is consistently higher when it is presented as an outlier item than when it is a non-outlier item in a ranking. Next,

to examine whether these findings can be generalized to naturalistic situations we perform an analysis on real-world search logs **from Bol.com, a popular Europe-based e-commerce platform.** The results confirm the findings of our user study. In addition, we observe that, on average, outlier items receive significantly more clicks than non-outlier items in the same lists. Moreover, users tend to interact more with lists that contain at least one outlier.

Outlier bias vs. context bias. We find that outlier bias affects user clicks such that users are more likely to interact with items that are presented as outliers, as well as their neighboring items. The closest concept to outlier bias is context bias in news-feed recommendation [44]. In the presence of context bias CTR is lower for items when surrounded by at least one very similar item than when they are surrounded by non-similar items. This is different from outlier bias, which emphasizes the *difference* between the outlier and the rest of the list. Moreover, observability is a key factor in detecting outliers in ranking as defined by [34]; this is not the case in context bias.

Accounting for outliers. Based on the findings of our user study and log analysis, we conclude that one should account for the effect of outliers when unbiasing user clicks for ULTR. To this end, we propose a click model, based on the examination hypothesis, called *outlier-aware position-based model* (OPBM), which accounts for both outlier and position bias. OPBM assumes the probability of a click depends on (i) examination, (ii) relevance, and (iii) the outlier's position (if it exists). We use regression-based expectation maximization to estimate the click propensities based on our proposed click model, OPBM. We verify the effectiveness of our outlier-aware model for estimating propensities in the presence of both position bias and outlier bias. Following [6, 19, 26] we use a semi-synthetic setup for the experiments; the true relevance labels provided in this setup allows for evaluating the relevance estimation. Furthermore, using simulated clicks we are able to control the severity of position bias and outlier bias. The results of our experiments show the superiority of OPBM against baselines in terms of ranking performance (NDCG@10) and true relevance estimation.

Main contributions. The main contributions of this work are: (i) we identify and study a new type of click bias, originating from inter-item dependencies, called outlier bias; (ii) through extensive analyses of both user study results and real-world search logs, we confirm our hypothesis about the existence of outlier bias; (iii) to address this effect we propose an outlier-aware click model that accounts for outlier items (if they exist), as well as position bias; (iv) using an empirical analysis based on real-world data and semi-synthetic experiments we show the effectiveness of our outlier-aware model in estimating click propensities; and (v) we make the data from our user study plus the code that implements our baselines and OPBM publicly available.

2 OUTLIERS IN RANKING

Outliers in ranking are items that observably stand out among the window of items that are presented to a user at once. We use the following definitions from [34] to introduce so-called outliers:

Definition 2.1 (Observable feature). An observable item feature, $\mathcal{F}$, is a characteristic of an item **in a list** that **can be** purely presentational in nature (e.g., image, title font size, and discount **tag**).

Definition 2.2 (Degree of outlierness). Let $\mathcal{M}$ be any outlier detection method, and $\mathcal{F}_i$ an observable feature corresponding to item i, in the context of all items in the list, $\mathcal{C}$. The degree of outlierness for item i is the value calculated by $\mathcal{M}$ for $\mathcal{F}_i$ w.r.t. $\mathcal{C}$ shown as $\mathcal{M}(\mathcal{F}_i|\mathcal{C})$. This value indicates how much the corresponding item differs from the other elements of the set **w.r.t.** $\mathcal{F}$.

Definition 2.3 (Outliers in ranking). Let $\mathcal{M}$ be any outlier detection method; we call item i in a ranked list an outlier, if $\mathcal{M}$ identifies it as an outlier w.r.t. an observable feature, $\mathcal{F}$, based on the degree of outlierness, and in the context of the list.

In Section 3.2 we describe our choices of observable features and outlier detection method used in this paper.

3 IMPACT OF ITEM OUTLIERNESS ON CLICKS

Sarvi et al. [34] show that outlier items receive more attention from users. However, it is not known whether an item's outlierness affects users' clicks as well. In this section we answer our first research question: (**RQ1**) does outlier bias exist in rankings of items? To this end we first conduct a user study to examine the outlierness effect as the only variable factor influencing the clicks. Next, we need to examine whether the findings of our study can be generalized to naturalistic situations. In other words, we seek to establish ecological validity [8, 21]. To this end, in Section 3.2 we explore real-world click logs to confirm our findings.

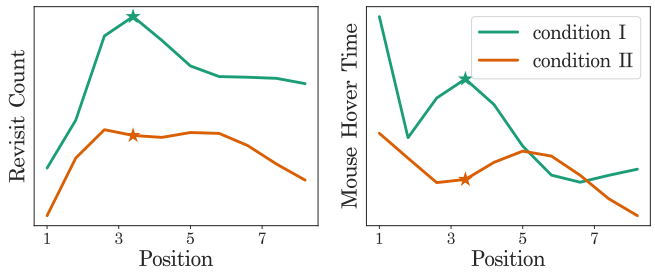
3.1 User study

In this section, we present the results of our user study. Our main goal is to learn whether the outlierness of an item affects user clicks, independent of the item's relevance and position.

Setup. **We mimic Bol.com, a popular European online marketplace.** We ask participants to interact with search engine result pages as they normally would, and find items they prefer and think are relevant. We focus on a list view, with 20 items on each page, and participants are able to scroll the list to see all items. We have two queries; for each query, we show one specific item once as an outlier and once as a regular item. We call this specific item the *target*, and these two variant presentations *condition I* and *condition II*, respectively. In condition I the target is an outlier w.r.t. a set of observable features, such as item category,[1] price, discount tag, and star rating. We aim to compare users' behavior between these two conditions for each query. We keep other factors such as relevance and position bias unchanged between the conditions. To eliminate the effect of position bias we always show the item at rank 4, and to maintain the same degree of relevance to the query we only change the surrounding items to change the outlierness of the target item.[2]

We also have a Qualtrics [31] survey. It contains the task instruction, multiple choice questions about the instructions, queries, and links to the examples, and a few demographic questions at the end. In the instructions, we describe the overall goal of the research and ask participants to read the instructions carefully. We describe what

[1]Note that this feature can affect the outlierness w.r.t. the item's image as well.
[2]To examine our hypothesis about an inter-item dependency, here we assume that the relevance of a document is only dependent on the query.



(a) Revisit count (b) Mouse hovering time

Figure 1: Revisit count (a) and mouse hovering time (b) for the two conditions of one of our user study examples. The position of the target item is marked by an asterisk. The plots show that the user engagement with the target item is higher when presented as an outlier (condition I)

it means to interact with a result page in terms of exploring the results, scrolling the list, and clicking on items that seem interesting. Participants can click on an item to open the item's detail page. In our instructions we encourage participants to click on items they find interesting, however, clicking is not mandatory. We instruct participants to first read and understand the query, and then scan the result page as if they submitted the query themselves.

Participants. We recruit 40 workers, based in countries where our marketplace is active, from the Prolific platform [29]. From the participants, 14 are female, 23 are male, and 3 listed other genders. The majority of participants (27) are between 25 and 44 years old, with 10 participants younger and 3 older; 33 participants reported that they shop online at least once a month.

Metrics. For reporting we consider three measures based on participants' interactions with rankings: (i) revisit count, which indicates how many times on average participants viewed an item (due to scrolling), (ii) mouse hover time that shows the amount of time on average participants spent on an item, and (iii) CTR for the target item in each condition, which is our main metric in this study.

Findings. We expect to see more interactions with the target in condition I. Since we keep other factors unchanged between the two conditions, we can attribute any difference in user behavior to the inter-item dependencies.

Figure 1 depicts the revisit counts (Figure 1a) and mouse hovering time (Figure 1b) for different positions and conditions of one example. Both plots show that the user engagement with the target item is higher when it is presented as an outlier. We see the same pattern in the second example. On average, participants revisited the outlier item more often and spent more time examining it. These findings are in line with the results of the eye-tracking experiments conducted by Sarvi et al. [34], which suggest that, on average, outlier items receive more attention from users. However, our main goal is to study if this increased attention leads to more clicks.

Table 1 reports the CTR for the target item in both examples and for the two conditions. In both examples we see a large difference between the CTR reported for the different conditions, suggesting that when the target item is shown as an outlier it receives more clicks as well as more exposure.

Table 1: CTR of the target item's position in both examples of our user study. The target item recieves more clicks when shown as an outlier (condition I).

	condition I	condition II
Query 1	0.944	0.166
Query 2	0.880	0.091

3.2 Real-world click logs

The findings of Section 3.1 confirm, in a controlled experimental setup, that an item's outlierness can influence users' click behavior. However, we still need to examine the ecological validity of this hypothesis. To this end, we present our observations of click exploration of **real-world search logs from our e-commerce platform**. We are specifically interested in exploring the data to study the existence of outlier items in rankings and their impact on click data. Notice that we use this data only for click analysis and parameter estimation (Section 5).

Data collection. We collect search query logs from 20 consecutive days. Each row of the dataset consists of seven observable item features that are explained in Table 2, along with users' interaction signals: impressions and clicks.

Definition 3.1 (Impression). An *impression* indicates how many times an item that is rendered by the search engine is viewed by a user. If an item is rendered in low positions, it may not end up in a window that is visible to the user, leading to zero impressions. On the other hand, the number of impressions can be greater than one due to scrolling.

We selected item features that are used across different categories, are observable by users, and have been shown by previous work to be important in influencing users' purchase decisions [4, 20]. We leave out item images from our click exploration due to the excessive complexity they would have added to this study.

Most search engines consider diversity as a quality of search result pages [5]. This can have a side effect, where the returned rankings may contain outlier items. Hence, query logs are a valuable source for studying the outliers' effect on users' clicking behavior. To begin, we define two types of rankings based on the existence of outliers as follow:

Definition 3.2 (Normal rankings). We call rankings that contain no outlier *normal* rankings. Normal rankings can either consist of a homogeneous set of items or a diverse set.

Definition 3.3 (Abnormal rankings). We define *abnormal* rankings to be lists that contain at least one outlier.

Outlier detection. We examine each item for outlierness based on the features described in Table 2 and in the context of all items in the list as described in Section 2. An item is an outlier if it is detected as an outlier w.r.t. at least one of these features.

We use the Interquartile rule to detect the outliers, and consider the absolute difference between the feature value and the upper/lower bound as the degree of outlierness of the corresponding item (see Section 2). Feature values are normalized so that we have an outlierness degree of unit range for all observable features. We set the threshold for the degree of outlierness to 0.5, which means we only label an item as an outlier if the absolute difference

Table 2: Description of the observable features used to represent the items.

Feature Name	Abbreviation	Description
price	-	Selling price of an item.
promotion tag	promotion	Universal red tag indicating various promotions, such as 'competitive price' and 'select deal'.
high discount tag	discount	Two-piece red tag indicating high discount for an item (different from promotion).
in/out-of-stock tag	stock quantity	Green tag indicating the in-stock or out-of-stock condition of an item.
users star rating	rating	Average user star rating of the item presented by the standard 5 stars template.
'select' tag	select	Green tag indicating that the item is a select item (similar to Amazon prime).
title length	-	Number of tokens in the item title.

Table 3: Users' interactions with the outlier and non-outlier items, averaged over all abnormal rankings. We used Student's t-test with $p < 0.001$ for statistical significance test.

	Avg. clicks	Avg. impressions	Avg. CTR
Outliers	0.202*	1.381*	0.142*
Non-outliers	0.137	1.346	0.098
Total	0.149	1.352	0.106

between its score and upper/lower bound is greater than 0.5.

Post-processing. We filter out the parts of the rankings that are not viewed by the user based on the impression signal in our data. This leaves us with the minimum ranking size of 3. However, since by definition outlierness is meaningless in lists shorter than 4, we removed these rankings from our dataset. We also removed pages with sponsored items to avoid any potential effect from such items on our results. The remaining 10,903 abnormal rankings have an average length of 10.24 and a median of 8.0.

Effect of outliers on CTR. In the first step of our analysis, we aim to see if users interact differently with outlier items in abnormal rankings. To this end, we look at such rankings and compare the number of interactions outliers received on average to non-outlier items in the same ranking. We focus on clicks as interactions.

Since normal rankings carry no information for our current analysis we only keep abnormal rankings. Table 3 shows the average clicks, impressions, and CTR of outlier and non-outlier items for abnormal rankings.[3] We calculate the CTR values (i.e., the number of clicks divided by the number of impressions of each item) per page and report the average over all rankings. Our findings suggest that both CTR and average clicks are significantly higher for outlier items when compared to non-outlier items on the same page. Moreover, we see that the number of impressions is also significantly higher for outlier items, which is in line with the finding of an eye-tracking experiment reported in [34].

Effect of outliers per position. To make sure that the higher CTR reported in Table 3 is not caused by position bias, we look at CTR values per position. Figure 2 depicts the results. Overall, CTR for all positions is higher for outlier items, showing that these items receive more interactions than non-outlier items.

Next, to further study how outliers change users' click behavior, we compare the CTR of the outlier position with the positions of non-outlier items throughout the ranking. To better depict the effect of outliers on different positions, we consider rankings that contain exactly one outlier; we focus on the top 15 positions. It is worth mentioning that less than 35% of the abnormal rankings in our data have more than one outlier. We group the abnormal rankings based on the position of the outlier. Figure 3 illustrates the results. The black line shows CTR for normal rankings. As expected this line follows position bias, where the probability of clicking an item decreases with its rank.

The other lines in Figure 3 show the CTR for groups of rankings with one outlier at position $r \in \{1, \ldots, 15\}$. We only show the results for some of the positions for better visibility. We see similar patterns for other ranks. We only show the results for groups that

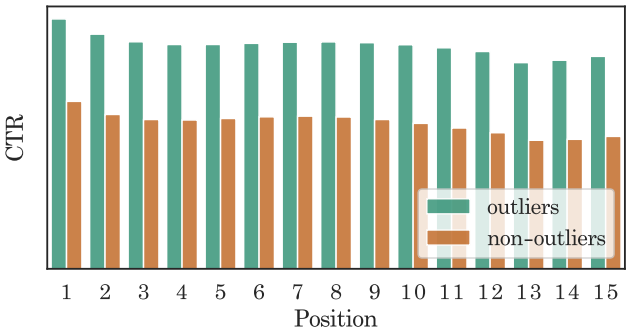


Figure 2: Comparison of CTR for outlier and non-outlier items per rank. CTR is consistently higher for outlier items.

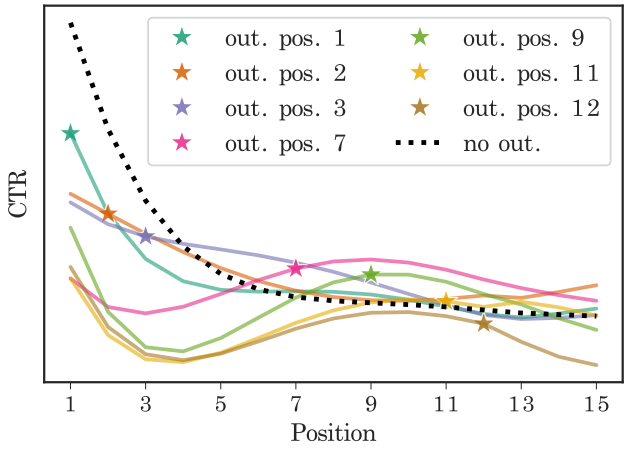


Figure 3: CTR per rank for abnormal rankings grouped by the outliers' position. The position of the outlier is marked with an asterisk. The values are smoothed using a Savitzky-Golay filter. Best viewed in color.

[3]Note that the reported values in this section are calculated based on filtered subsets of search logs, therefore, they are not representative of the true statistics of the data.

form at least 1% of the whole collection in terms of size. In Figure 3 each asterisk indicates the position of the outlier. We observe that CTR distribution is different than the position-based assumption when there is an outlier in the ranking. Also, for positions after 3, we observe an increase of CTR on and around the outlier position.

Another interesting observation is that items farther away from the outlier receive less attention proportional to their ranks compared to normal pages. Moreover, we see that for positions after 3 CTR for outliers is higher than the CTR for the same position in normal pages, which is in line with our findings in Figure 2.

Effect of different outlier types. One can argue that different types of outliers might have different types of influence on users' perceptions. E.g., considering price as the observable feature, a very expensive outlier item might have a lower chance of being purchased compared to a cheap one. We hypothesize that these two types may neutralize each other overall in terms of statistical metrics. Hence, to examine this hypothesis, as a first attempt, we divide the outlier items into two groups of positive and negative outliers using common sense, informal definitions based on the observable features. E.g., in the previous example, the expensive item is a negative outlier while the cheap one is positive. Based on this definition, for the price feature we see that the average number of clicks for positive and negative outliers are 0.193 and 0.147, respectively; both are significantly higher than non-outlier items (0.125). We see the same trend among all observable features, both for impression and click counts. Based on these results we reject the aforementioned hypothesis and stay with our original outlier/non-outlier division.

Further remarks. We also looked at abnormal rankings in which a specific item is repeatedly shown in a fixed rank at least once as an outlier and once as a normal item. We aggregate all such rankings and observe that on average items receive 0.169 clicks in case of being an outlier, and 0.130 clicks when they are regular items in the list. Comparing the abnormal rankings to a subset of normal rankings with a similar length distribution (mean=10.09/10.30, median=8.0/8.0, std=4.95/5.82 for normal/abnormal rankings), we realize that on average the number of clicks per session is higher in the presence of outliers. More specifically, the average number of clicks is 0.139 for normal rankings, and 0.149 for abnormal rankings, with a $p < 0.001$ significance.

Upshot. To sum up, from Section 3.1 we learn that users behave differently w.r.t. an item given its outlierness condition (i.e., whether the item is presented as an outlier in the ranking). The CTR of a specific item is consistently higher when it is presented as an outlier item than that of a non-outlier item. Section 3.2 confirms the findings of our user study. In addition, we observe that, on average, outlier items receive significantly more clicks than non-outlier items on the same lists. Moreover, users tend to interact more with lists that contain at least one outlier. This section confirms the impact from outlier items on clicks. We refer to this effect as outlier bias. In the following section we propose a click model that accounts for outlier bias as well as position bias.

4 OUTLIER-AWARE POSITION-BASED MODEL

Naïve use of implicit feedback for learning to rank can be misleading, since it suffers from presentation bias. Therefore, modeling the examination bias is crucial [13, 18].

Position-based model. Normally, items in higher ranks are more likely to be examined on a page. Position bias is formally modeled through the examination hypothesis which states that an item must be examined and perceived relevant by the user to be clicked. A widely used click model for dealing with position bias is the position-based model (PBM) [19, 43]. While being considered a simple solution, PBM is as effective as more sophisticated click models [12]. PBM assumes that the rank of an item is the only parameter that affects users' examination of that item. Examining an item means viewing and evaluating it before any subsequent interaction like a click.

Given an item d at rank k in response to a query q, the probability of clicking on d, assuming PBM, equals:

$$P(C = 1 \mid q, d, k) = P(E = 1 \mid k) \times P(R = 1 \mid d, q), \quad (1)$$

where $P(E = 1 \mid k)$ is the probability of user examining rank k, also called *propensity*, and $P(R = 1 \mid d, q)$ is the probability of relevance for the pair (d, q). We refer to these probabilities as θ_k and $\gamma_{q,d}$, respectively.

Outlier-aware position-based model. PBM simply assumes that the only factor influencing the propensity is the rank. In Section 3 we show that users are more likely to click on outlier items, hence, we assume that propensity depends also on the existence of outlier item(s).

It is noteworthy that, even among the outlier items we observe an inter-outlier position bias – the higher-ranked outlier items receive more clicks.

Hence, to model these dependencies, we propose an outlier-aware position-based model, called OPBM, that accounts for the impact of outlier items in addition to the position as follows:

$$P(C = 1 \mid q, d, k, o) = P(E = 1 \mid k, o) \times P(R = 1 \mid d, q), \quad (2)$$

where o indicates the **position(s)** of the outlier(s) in the ranking. Note that PBM is a special case of OPBM: for normal rankings OPBM is simplified to PBM.

We propose this model following Eq. (2) based on the assumption that the probability of examination at rank k depends on the position of outlier item(s), o, in addition to k. This model has $K \times O$ parameters, where K and O are the set of all ranks and outlier positions, respectively, which can be estimated from click data.

Propensity estimation. Here, we describe how to estimate outlier-aware position bias from regular clicks. Based on the idea of the regression-based expectation maximization (REM) algorithm [43], we propose to estimate the parameters $\theta_{k,o}$ and $\gamma_{q,d}$ simultaneously by estimating with a regression function.

Using a standard expectation maximization (EM) algorithm we aim to find the parameters that maximize the log-likelihood of the whole click logs. The log likelihood of generating click logs of the form $\mathcal{L} = (c, q, d, k, o)$ is:

$$\log P(\mathcal{L}) = \sum_{(c,q,d,k,o) \in \mathcal{L}} c \log \theta_{k,o}\gamma_{q,d} + (1-c) \log(1 - \theta_{k,o}\gamma_{q,d}). \quad (3)$$

Here, we aim to estimate the parameters $\theta_{k,o}$ and $\gamma_{q,d}$ based on data points in $\mathcal{L}$. In each iteration, EM alternates between the expectation and maximization steps to compute new estimates of the parameters. In the expectation step of iteration $t+1$ we calculate the

hidden variables corresponding to examination propensity (E) and true relevance (R) based on the estimated parameters at iteration t:

$$
\begin{aligned}
P(E=1,R=1 \mid C=1,q,d,k,o) &= 1,\\
P(E=1,R=0 \mid C=0,q,d,k,o) &= \frac{\theta^t_{k,o}(1-\gamma^t_{q,d})}{1-\theta^t_{k,o}\gamma^t_{q,d}},\\
P(E=0,R=1 \mid C=0,q,d,k,o) &= \frac{(1-\theta^t_{k,o})\gamma^t_{q,d}}{1-\theta^t_{k,o}\gamma^t_{q,d}},\\
P(E=0,R=0 \mid C=0,q,d,k,o) &= \frac{(1-\theta^t_{k,o})(1-\gamma^t_{q,d})}{1-\theta^t_{k,o}\gamma^t_{q,d}}.
\end{aligned} \tag{4}
$$

We then calculate the marginal probabilities $P(E=1 \mid c,q,d,k,o)$ and $P(R=1 \mid c,q,d,k)$ for each data point in $\mathcal{L}$. We keep the estimation of $\gamma_{q,d}$ untouched, meaning that the learning to rank (LTR) model is trained without knowledge of the outlier position and only the propensity estimation is affected by that. This leads to the maximization step at iteration $t+1$, where we update the parameters to maximize the likelihood from Eq. 3 as follows:

$$
\begin{aligned}
\theta^{t+1}_{k,o} &= \frac{\sum_{c,q,d,k',o'} \mathbb{I}_{k'=k,o'=o}.(c+(1-c)P(E=1 \mid c,q,d,k,o))}{\sum_{c,q,d,k',o'} \mathbb{I}_{k'=k,o'=o}},\\
\gamma^{t+1}_{q,d} &= \frac{\sum_{c,q',d',k} \mathbb{I}_{q'=q,d'=d}.(c+(1-c)P(R=1 \mid c,q,d,k))}{\sum_{c,q',d',k} \mathbb{I}_{q'=q,d'=d}}.
\end{aligned} \tag{5}
$$

The maximization step of the EM algorithm requires multiple occurrences of pair (q,d) where d is shown in different positions. To overcome the click sparsity problem and possible privacy issues, we alter the maximization step at iteration $t+1$, where we estimate the $\gamma_{q,d}$ parameter via regression [43]. Thus, given a feature vector $x_{q,d}$ representing the pair (q,d) we fit a function $f(x_{q,d})$ (e.g., gradient boosted decision tree (GBDT)) to calculate an estimate for $\gamma_{q,d}$. So, our maximization step is to find a regression function $f(x)$ that maximizes Eq. 3 given the estimated parameters from the expectation step. In **REM algorithm** [43], this regression problem is converted to a classification problem by sampling a binary variable indicating the relevance label for $x_{q,d}$ from the distribution $P(R=1 \mid c,q,d,k)$. This results in a training set of the form $(x_{q,d}, r_{q,d})$ with the following cross entropy objective:

$$
\sum_{x,r} r\log(f(x)) + (1-r)\log(1-f(x)). \tag{6}
$$

Remark. An alternative choice instead of a single unbiased model would be to train multiple LTR models as unbiased experts for different outlier positions. This alternative has two main drawbacks. First, having experts means that each expert is trained only on a part of the data containing outliers at a specific position. Not only can this lead to sub-optimal training, but it also makes it difficult to compare this model to the PBM-based REM as a baseline. Second, having a collection of K expert models as a ranker is not ideal in real-world scenarios. Ideally, there is a single unbiased model that can be used without information about outliers' positions.

5 EXPERIMENTAL SETUP

Following much previous work in counterfactual learning to rank (CLTR) [6, 15, 19, 26, 38], we use a semi-synthetic setup for our experiments, i.e., we sample queries, documents, and relevance labels from existing LTR datasets, but simulate user clicks based on the probabilistic click models estimated on the proprietary data.

LTR datasets that contain the true relevance labels allow us to evaluate the relevance estimation of OPBM and other baselines, as well as their effect on ranking performance. Furthermore, the semi-synthetic setup enables us to control the position bias and outlier bias of the simulated clicks.

5.1 Data

Public LTR data. Following prior work on CLTR [19, 37, 38], we use the Yahoo! Webscope [10] and MSLR-WEB30k [30] datasets. In both datasets, there are a total of around 30k queries, each associated with a list of documents. The query-document feature vectors of the Yahoo! and MSLR datasets have dimensions 501 and 131, respectively. Both datasets have graded relevance labels with 5 levels. We follow prior work and take grades $\{3,4\}$ as relevant and grades $\{0,1,2\}$ as non-relevant. The training sets of the Yahoo! and MSLR datasets have 20k and 19k queries with 473k and 2.2M documents, respectively. The test sets of the Yahoo! and MSLR datasets, have 6.7k and 6k queries with 163k and 749k documents, respectively.

Proprietary data. We use the real-world click log data as described in Section 3.2 for the experiments and refer to it as *proprietary* data. **We** use a feature vector of size 24 containing both the relevance features and products' observable features to present each query-document pair. We use these features for the LTR model. We also use the setup described in Section 3.2 to detect the outlier items, using the Interquartile rule, w.r.t. the observable features (see Table 2). **Since the rankings in this dataset** have an average length of 10.24 and a median of 8.0, we use the top-10 items in the experiments.

5.2 Click simulation

We follow prior work [6, 19, 26, 37, 38] and sample 1% of the queries from each public dataset, uniformly at random, to train an artificial production ranker. We apply probabilistic click models on rankings produced by this production ranker to simulate clicks for the semi-synthetic experiments. We apply our outlier-aware position-based model with different approximations for examination probabilities. The relevances $\gamma_{q,d}$ are based on the relevance label recorded in the datasets. Following previous work [19, 38] we use binary relevance:

$$
P(R=1 \mid q,d) = \gamma_{q,d} = \begin{cases} 1 & \text{if relevance\_label}(q,d) > 2 \\ 0 & \text{otherwise} \end{cases}. \tag{7}
$$

To simulate the outlier bias we follow two strategies **as follows:**

OPBM$_{Real}$. We use the propensities estimated by OPBM (see Section 4) on our proprietary dataset. From all the abnormal rankings in our dataset, 64% contain only one outlier. Since improving ranking for more than half of queries can lead to significant improvement in real-world scenarios, we first address this majority case. Therefore, with this model, we focus on rankings with one outlier. Thus, the output of OPBM is at most a $K \times K$ matrix, corresponding to all combinations of rank and outlier position, where $K=10$ in our experiments. We use this matrix to approximate $P(E=1 \mid k,o)$.

OPBM$_{\mathcal{G}}$. Here, we assume that an outlier's effect on the user clicks follows a Gaussian distribution, centered at the outlier's position. Therefore, for each k, we compute the linear interpolation of outlier

bias, and position bias distributions, as follows:

$$OPBM_{\mathcal{G}}(q, d, k, o) = \gamma_{q,d}((1-\alpha)\theta_k + \alpha\mathcal{G}(\mu = o, \sigma^2)), \quad (8)$$

where $\mathcal{G}$ is a Gaussian distribution with $\mu = o$, simulating the outlier effect. We set $\sigma = 1$ and experiment with varying values of α. To simulate clicks for rankings with multiple outliers, we compute the average of OPBM$_{\mathcal{G}}$ for all outlier positions (O') as follows:

$$OPBM_{\mathcal{MG}}(q, d, k, O') = \frac{1}{|O'|}\sum_{i \in O'} OPBM_{\mathcal{G}}(q, d, k, i). \quad (9)$$

According to our proprietary data, 91% of abnormal rankings contain at most two outliers. Therefore, in the experiments, we focus on rankings with a maximum of two outliers.

We follow previous work [15, 19, 26, 38] to define the position bias inversely proportional to the item's rank as:

$$\theta_k = \frac{1}{k}. \quad (10)$$

We train the LTR model[4] on 1M simulated clicks.

5.3 Methods used for comparison

Our main goal is to introduce a new type of bias and study its impact on click propensities. Hence, it suffices to compare our outlier-aware click model to baselines that only corrects for position bias. To this end, we compare OPBM with the following estimators:

- **Naïve** is a model with no correction where each click is treated as an unbiased relevance signal.
- **PBM** is the original inverse propensity scoring (IPS) estimator [19, 43] that only corrects for position bias.

5.4 Evaluation metrics

To measure the ranking performance achieved by different methods we use normalized discounted cumulative gain (NDCG). We also consider cross entropy (CE), which measures the difference between the true relevance and unbiased relevance calculated by the estimator; it is an indication of how accurately a model estimates the relevance, independent of the LTR model. Since we work with binary relevance, we compute binary CE between the corrected clicks, i.e., c/θ_k and $c/\theta_{k,o}$ for PBM and OPBM, respectively, as predictions and the true relevance values as labels. We report the mean value of CE instead of its summation, for better readability.

6 RESULTS

In Section 3 we have already answered (**RQ1**) about the existence of outlier bias in ranked lists. In this section we answer the following research questions: (**RQ2**) how does our outlier-aware model, OPBM, perform compared to the baselines? (**RQ3**) how does OPBM perform under different outlier bias severity conditions? (**RQ4**) how does OPBM generalize to cases with multiple outliers in rankings?

6.1 Propensity estimation with OPBM

We answer (**RQ2**) by comparing the overall performance of OPBM in propensity estimation. Figure 4 depicts the propensities estimated by OPBM$_{Real}$ (see Section 5.2) on the top-8 ranks where a sufficient number of outliers exist in our proprietary dataset. We see that

[4]We use allRank implementation for our LTR [28].

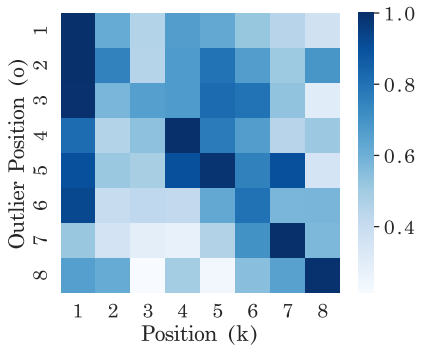


Figure 4: Click propensities computed by OPBM for the top 8 ranks, per outlier position on the proprietary data. Click propensities are higher on and around the outliers, contradicting the position bias assumption.

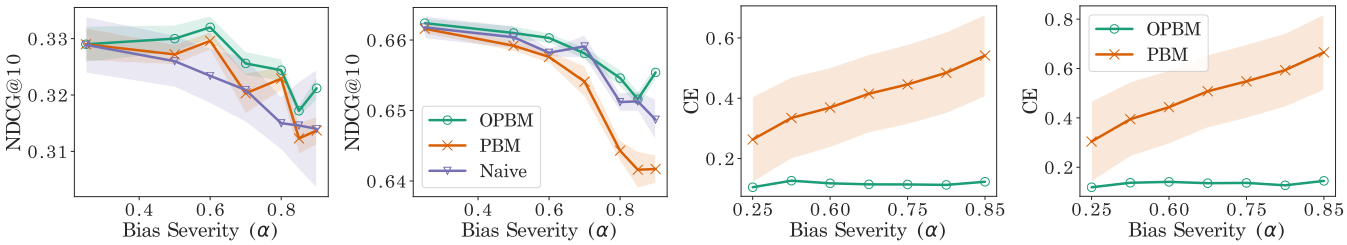
Table 4: Comparison of OPBM and PBM on the Yahoo! and MSLR datasets, in terms of NDCG@10 and CE. A superscript $*$ indicates a significant difference compared to the second-best performing method with $p < 0.001$.

	MSLR		Yahoo!	
	CE↓	NDCG@10↑	CE↓	NDCG@10↑
Oracle	-	0.3451	-	0.6713
Naïve	0.8205	0.3065	0.9786	0.6489
PBM	0.5474	0.3165	0.6807	0.6406
OPBM	0.1732*	0.3233*	0.1916*	0.6470*

the propensities are highest on and around the outlier positions which is in line with our findings in Section 3. However, this effect is less evident in the top-3 ranks. This is expected since we observe that position bias dominates in the top-3 ranks (see Section 3.2), diminishing the effect of outliers. Nevertheless, the effect of position bias decreases as the outlier appears higher in the ranking. For example, when an outlier occurs at position 1, the propensities of the first two ranks are 0.99 and 0.62, respectively. However, when the outlier occurs at position 7, these values reduce to 0.52 and 0.35.

Next, we report the results of the semi-synthetic experiments. We use the MSLR and Yahoo! public LTR datasets with simulated clicks. We use the propensities calculated by OPBM$_{Real}$ trained on our proprietary data. We compare OPBM and PBM in terms of relevance estimation (CE) and ranking performance (NDCG@10). Table 4 summarizes the results; on both datasets OPBM performs significantly better than PBM in terms of CE ($p < 0.001$), indicating that OPBM approximates click propensities more effectively – it estimates true relevance of a (q, d) pair more accurately. Providing an accurate estimate of true relevance is crucial in domains such as exposure-based fair ranking [9, 24, 36], where relevance is used as an indication of an item's merit [9, 14, 24, 34, 36, 39], and can have a big impact on fairness estimation. Table 4 also shows that OPBM significantly improves the ranking scores (NDCG@10) over the PBM baseline, again on both datasets.

In conclusion, using OPBM leads to more accurate propensity estimations and a more accurate approximation of the true relevance in rankings affected by outlier items. We also observe significant improvements in ranking performance by OPBM over PBM on the Yahoo! and MSLR datasets.



(a) MSLR (b) Yahoo! (c) MSLR (d) Yahoo!

Figure 5: Comparison of different estimators in term of NDCG@10 ((a) MSLR and (b) Yahoo!) and CE ((c) MSLR and (d) Yahoo!) under varying levels of outlier bias. Results are averaged over 8 runs; shaded area indicates the standard deviation.

Table 5: Comparison of OPBM, OPBM$_{lazy}$ and OPM on the Yahoo! and MSLR datasets, with outlier bias severity of $\alpha = 0.75$, and in terms of NDCG@10 and CE. A superscript $*$ indicates a significant difference with PBM with $p < 0.001$.

	MSLR		Yahoo!	
	CE↓	NDCG@10↑	CE↓	NDCG@10↑
Naïve	0.5704	0.3159	0.6776	0.6564
PBM	0.3126	0.3219	0.3958	0.6497
OPBM$_{lazy}$	0.1374*	0.3223	0.1548*	0.6566*
OPBM	0.1283*	0.3229	0.1407*	0.6572*

6.2 Effect of outlier bias severity

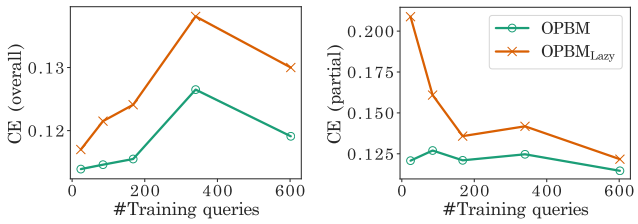
Next, we address (RQ3) by considering the impact of outlier bias severity on the performance of OPBM. For the sake of simplicity, we assume that outliers have the same effect on propensity distribution independent of their position; we use OPBM$_{\mathcal{G}}$ (see Section 5.2) for click simulation. The parameter α in OPBM$_{\mathcal{G}}$ allows us to control outlier bias severity. Figure 5 depicts the results. OPBM consistently outperforms PBM in terms of ranking performance. The results on Yahoo! dataset (Figure 5b) clearly show that the difference in ranking performance of the two models increases with the severity of outlier bias. In the case of MSLR (Figure 5a) we observe more fluctuations in OPBM's performance. This is also visible in the high variance of Naïve's performance in different runs; OPBM performs more robust compared than Naïve and PBM. Moreover, the results show that OPBM performs similarly to PBM at its worst, making it a more reliable choice as a user examination model for all severity levels of outlier bias. This is in line with our theory, which indicates that PBM is a specific case of OPBM (see Section 4).

In terms of cross entropy (Figure 5c and 5d), OPBM consistently outperforms PBM with a high margin. Also, the high variance in performance of PBM emphasizes the much more robust performance of OPBM compared to PBM in relevance estimation.

In conclusion, using the OPBM estimator leads to improved ranking models compared to PBM, especially when severe outlier bias exists. This finding also holds for accurately estimating the true relevance scores (i.e., CE). In the presence of slight outlier bias, OPBM exhibits a similar performance compared to PBM, making it a natural choice as it proves to be reliable.

6.3 Generalization to multiple outliers

We address (RQ4) by considering how OPBM generalizes to multiple outliers in the ranking. For click simulation, we use Equation 9



(a) Over all rankings (b) Over two outlier rankings

Figure 6: Comparison of OPBM and OPBM $_{lazy}$ on varying sizes of queries with multiple outliers.

with severe outlier bias ($\alpha = 0.75$). As pointed out before, our proprietary data shows that 91% of abnormal rankings contain at most two outliers. Therefore, we report results for $|O'| = 2$. Here, in addition to the single outlier rankings from the previous experiments, our semi-synthetic data contains rankings with two outliers at positions 4 and 9. As mentioned earlier, position bias is severe in the top-3 ranks, thus we place the first outlier in the fourth position of the list. Then, in order to see the effect of the outliers separately, we choose the second positions with some distance (rank 9).

To model the effect of multiple outliers, we propose two strategies: (i) According to the original description of OPBM (see Section 4), we consider the condition of having multiple outliers as a separate value for o, i.e., we separately compute the click propensities for k ranks, when two outliers exist in the ranking at positions 4 and 9. (ii) We simplify the problem and only consider the first outlier position, and call it OPBM $_{lazy}$. We compare the performance of OPBM between these two strategies and also with PBM. Table 5 summarizes the results. Overall, we see that both variations of OPBM outperform PBM in terms of NDCG@10 and CE. As expected, OPBM outperform its simpler version, OPBM $_{lazy}$ w.r.t. both metrics; we see significant improvements in CE, while the improvements over ranking performance are marginal. We can conclude that the original version of OPBM as the exact solution performs better for cases with multiple outliers. However, in case of data sparsity we can reduce the problem to the single outlier setup and still achieve higher results than PBM.

Lastly, we provide insights into how the size of the training data influences the performance of OPBM compared to OPBM $_{lazy}$. We gradually increase the number of training queries for the rankings with two outliers (positions 4 and 9), while keeping the rest of the training set unchanged. We compare the performance in term of CE on all rankings (overall), and only on the two outlier rankings

(partial). See Figure 6. We see that even with 24 rankings with two outliers, OPBM manages to learn the propensities better than the OPBM $_{lazy}$. However, the difference between the total performance (Figure 6a) of the two models grows by the size of training samples for the two outliers rankings, suggesting OPBM as a natural choice when a reasonable amount of training data is available.

In conclusion, when enough samples corresponding to multiple outlier positions are available in the training data, it is best to use OPBM with a specific o that represents the case at hand. Otherwise, reducing these samples to the single outlier setting, by only considering the first outlier position, still outperforms PBM.

7 RELATED WORK

Outliers. An outlier is an exceptional object that deviates from the general data distribution [40]. Outliers can affect the statistical analysis, whether they are interesting observations or suspicious anomalies. Identifying these outlaying samples is crucial in many fields of study [22, 40]. Numerous approaches have been proposed to detect outliers [16, 22, 32, 33, 35, 47]. Defining and dealing with outliers is dependent on the application domain [40]. We follow the definition of outliers in ranking from [34]: outliers are items that stand out in the ranking w.r.t. observable item features. They study the effect of such items on the exposure distribution through eye-tracking experiments and further address the effect of outliers on exposure-based fairness. In contrast, in this work we focus on click bias caused by this phenomenon. We are the first to investigate the existence of outlier bias in real-world search click logs and to propose an ULTR model to correct for outlier and position bias.

Bias in implicit feedback. Users' implicit feedback, such as clicks, can be a valuable source of supervision for CLTR [2]. However, the bias in click data can cause the probability of a click to differ from the probability of relevance, which is misleading. In recent years, different types of bias have been studied, such as position [17, 19], presentation [46], selection [27], trust [2, 38], popularity [1], and recency bias [11]. Another factor influencing the perceived relevance of items is inter-item dependency [12, 34]. We introduce outlier bias, which is a type of inter-item dependency. As outlier bias considers inter-item relationships it differs from the previously mentioned types of bias. Our work suggests that users tend to interact more with outlier items such that the examination probabilities assumed by position bias change when outlier items exist in the ranking.

Presentation bias [46] considers a related phenomenon; items with bold keywords in their titles appear more attractive. This differs from outlier bias by defining attractiveness of an item independent of its surrounding items. Moreover, adding more images to the top positions in a search result page can influence CTR [23]. However, the effect of such manipulations on click bias has not been studied. The closest concept to our work is context bias in news-feed recommendation [44]; CTR is lower for products when surrounded by at least one very similar product than when surrounded by non-similar products. This differs from outlier bias, which emphasizes the difference between the outlier and the other items. Also, observability is a key factor in detecting outliers [34], but context bias does not consider this factor. Unlike previous work, we focus on the effect of outliers on clicks, which is observable by users and comes from inter-item dependencies.

Unbiased learning to rank. Unbiased learning to rank approaches train an unbiased ranking model directly with biased user feedback [7]. These approaches can be classified into counterfactual learning to rank algorithms [6, 19, 42] and the bandit learning algorithm [25, 41, 45]. In this paper we are concerned with CLTR. The key factor in CLTR algorithms is first estimating examination probabilities [6, 43] and then using IPS [19, 42] to debias clicks. The estimations can be derived from online result randomization [42], online interleaving [19], or intervention data harvested from multiple rankers [3]. However,interventions can hurt user experience; Ai et al. [6] propose a dual learning algorithm to automatically learn both ranking models and propensities from offline data. Similarly, Wang et al. [43] use regression-based expectation maximization to compute the likelihood of observed clicks for each query. We build on [43] and propose an unbiased ranking model that corrects for both position bias and outlier bias by adding a parameter that accounts for the position of outlier(s).

8 CONCLUSION

We have introduced and studied a new type of click bias, that is, outlier bias. We conduct a user study to compare the CTR for specific items in two conditions: once shown as outliers and once as non-outlier items in the list. We find that the CTR is consistently higher when the item is presented as an outlier than when it is a non-outlier item. Moreover, our analysis on real-world search logs confirms the findings of our user study. On average, outlier items receive significantly more clicks than non-outlier items in the same lists.

To account for this effect, we propose OPBM, a click model based on the examination hypothesis, which accounts for both outlier and position bias. We use regression-based expectation maximization to estimate the click propensities based on our proposed click model, OPBM. Our experiments show (i) the superiority of OPBM against compared models in terms of ranking performance, and (ii) that true relevance estimation outlier bias exists. We show that OPBM performs more robustly on all levels of outlier bias severity compared to PBM. Moreover, our results show that OPBM performs similarly to PBM in the worst case, making it a more reliable choice.

One limitation of our work is that for rankings with multiple outliers, we assume that the effect of each outlier is independent of its position and other outliers. We plan to investigate how multiple outliers on the same ranking affect each other and their surrounding items. Finally, a natural extension of our work is to study how outlier bias can compensate for position bias in the top-k ranks, and explore its use in different domains such as fairness of exposure.

Data and code. To facilitate reproducibility of our work, all code and parameters are shared at https://github.com/arezooSarvi/outlierbias/.

ACKNOWLEDGMENTS

This research was supported by Ahold Delhaize and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-rank Recommendation. In *RecSys*. 42–46.

[2] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. 2019. Addressing Trust Bias for Unbiased Learning-to-rank. In *WWW*. 4–14.

[3] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. 2019. Estimating Position Bias without Intrusive Interventions. In *WSDM*. 474–482.

[4] Praveen Aggarwal and Rajiv Vaidyanathan. 2016. Is Font Size a Big Deal? A Transaction–Acquisition Utility Perspective on Comparative Price Promotions. *Journal of Consumer Marketing* (2016).

[5] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *WSDM*. 5–14.

[6] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. 2018. Unbiased Learning to Rank with Unbiased Propensity Estimation. In *SIGIR*. 385–394.

[7] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. 2021. Unbiased Learning to Rank: Online or Offline? *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–29.

[8] Chittaranjan Andrade. 2018. Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation. *Indian journal of psychological medicine* 40, 5 (2018), 498–499.

[9] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*. 405–414.

[10] Olivier Chapelle and Yi Chang. 2011. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research* 14 (2011), 1–24.

[11] Ruey-Cheng Chen, Qingyao Ai, Gaya Jayasinghe, and W Bruce Croft. 2019. Correcting for Recency Bias in Job Recommendation. In *CIKM*. 2185–2188.

[12] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers.

[13] Zhichong Fang, Aman Agarwal, and Thorsten Joachims. 2019. Intervention Harvesting for Context-dependent Examination-bias Estimation. In *SIGIR*. 825–834.

[14] Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. 2022. Fairness of Exposure in Light of Incomplete Exposure Estimation. In *SIGIR*. 759–769.

[15] Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke. 2019. To Model or to Intervene: A Comparison of Counterfactual and Online Learning to Rank from User Interactions. In *SIGIR*. 15–24.

[16] Wen Jin, Anthony K. H. Tung, Jiawei Han, and Wei Wang. 2006. Ranking Outliers Using Symmetric Neighborhood Relationship. In *PAKDD*. 577–593.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *SIGIR*. 154–161.

[18] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7–es.

[19] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-rank with Biased Feedback. In *WSDM*. 781–789.

[20] Karen C Kao, Sally Rao Hill, and Indrit Troshani. 2020. Effects of Cue Congruence and Perceived Cue Authenticity in Online Group Buying. *Internet Research* (2020).

[21] David J Lewkowicz. 2001. The Concept of Ecological Validity: What Are Its Limitations and Is It Bad to Be Invalid? *Infancy* 2, 4 (2001), 437–450.

[22] Zheng Li, Yue Zhao, N Botta, C Ionescu, and X COPOD Hu. 2020. COPOD: Copula-based Outlier Detection. In *ICDM*. 17–20.

[23] Pavel Metrikov, Fernando Diaz, Sebastien Lahaie, and Justin Rao. 2014. Whole Page Optimization: How Page Elements Interact with the Position Auction. In *EC*. 583–600.

[24] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-rank. In *SIGIR*. 429–438.

[25] Harrie Oosterhuis and Maarten de Rijke. 2018. Differentiable Unbiased Online Learning to Rank. In *CIKM*. 1293–1302.

[26] Harrie Oosterhuis and Maarten de Rijke. 2020. Policy-aware Unbiased Learning to Rank for Top-k Rankings. In *SIGIR*. 489–498.

[27] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-rank Systems. In *WWW*. 1863–1873.

[28] Przemyslaw Pobrotyn, Tomasz Bartczak, Mikolaj Synowiec, Radoslaw Bialobrzeski, and Jaroslaw Bojar. 2020. Context-Aware Learning to Rank with Self-Attention. *ArXiv* abs/2005.10084 (2020).

[29] Prolific. 2023. Data Annotation. https://www.prolific.co/.

[30] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597* (2013).

[31] Qualtrics. 2023. XM. https://www.qualtrics.com/.

[32] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient Algorithms for Mining Outliers from Large Data Sets. In *SIGMOD*. 427–438.

[33] Peter J Rousseeuw and Katrien Van Driessen. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41, 3 (1999), 212–223.

[34] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. In *WSDM*. 861–869.

[35] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the Support of a High-dimensional Distribution. *Neural Computation* 13, 7 (2001), 1443–1471.

[36] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*. 2219–2228.

[37] Ali Vardasbi, Maarten de Rijke, and Ilya Markov. 2021. Mixture-Based Correction for Position and Trust Bias in Counterfactual Learning to Rank. In *CIKM*. 1869–1878.

[38] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. 2020. When Inverse Propensity Scoring does not Work: Affine Corrections for Unbiased Learning to Rank. In *CIKM*. 1475–1484.

[39] Ali Vardasbi, Fatemeh Sarvi, and Maarten de Rijke. 2022. Probabilistic Permutation Graph Search: Black-Box Optimization for Fairness in Ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. 715–725.

[40] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. 2019. Progress in Outlier Detection Techniques: A Survey. *IEEE Access* 7 (2019), 107964–108000.

[41] Huazheng Wang, Ramsey Langley, Sonwoo Kim, Eric McCord-Snook, and Hongning Wang. 2018. Efficient Exploration of Gradient Space for Online Learning to Rank. In *SIGIR*. 145–154.

[42] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to Rank with Selection Bias in Personal Search. In *SIGIR*. 115–124.

[43] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *WSDM*. 610–618.

[44] Xinwei Wu, Hechang Chen, Jiashu Zhao, Li He, Dawei Yin, and Yi Chang. 2021. Unbiased learning to rank in feeds recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 490–498.

[45] Yisong Yue and Thorsten Joachims. 2009. Interactively Optimizing Information Retrieval Systems as A Dueling Bandits Problem. In *ICML*. 1201–1208.

[46] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *WWW*. 1011–1018.

[47] Yue Zhao, Zain Nasrullah, Maciej K Hryniewicki, and Zheng Li. 2019. LSCP: Locally Selective Combination in Parallel Outlier Ensembles. In *ICDM*. 585–593.