Online Evaluation Methods for the Causal Effect of Recommendations

MASAHIRO SATO*, Independent Researcher, Japan

Evaluating the causal effect of recommendations is an important objective because the causal effect on user interactions can directly leads to an increase in sales and user engagement. To select an optimal recommendation model, it is common to conduct A/B testing to compare model performance. However, A/B testing of causal effects requires a large number of users, making such experiments costly and risky. We therefore propose the first interleaving methods that can efficiently compare recommendation models in terms of causal effects. In contrast to conventional interleaving methods, we measure the outcomes of both items on an interleaved list and items not on the interleaved list, since the causal effect is the difference between outcomes with and without recommendations. To ensure that the evaluations are unbiased, we either select items with equal probability or weight the outcomes using inverse propensity scores. We then verify the unbiasedness and efficiency of online evaluation methods through simulated online experiments. The results indicate that our proposed methods are unbiased and that they have superior efficiency to A/B testing.

1 INTRODUCTION

A recommendation is a treatment that can affect user behavior. An increase in user actions, such as purchases or views, by the recommendation is the treatment effect (also called the causal effect). Because this leads to improved sales or user engagement, the causal effect of recommendations is important for businesses. While most recommendation methods aim for accurate predictions of user behaviors, there may be a discrepancy between the accuracy and the causal effect of recommendations [25]. Several recent works have thus proposed recommendation methods to rank items by the causal effect of recommendations [1, 24, 25, 27, 28].

Online experiments are commonly conducted to compare model performance and select the best recommendation model. However, evaluating the causal effect is not straightforward; we cannot naively compare the outcomes of recommended items because the causal effect is the difference between the *potential outcomes* with and without the treatment [12, 22]. A/B testing that compares the total user actions on all items, not only recommended items, can reveal the difference in the average causal effect (see Section 3.2). Nevertheless, it suffers from large fluctuations due to the variability in natural user behaviors for non-recommended items: some users tend to purchase more items than others. A large number of users is required to compensate for such fluctuations, making online experiments costly and risky.

In this paper, we propose efficient online evaluation methods for the causal effect of recommendations based on interleaving. Interleaving generates a list from the lists ranked by the two models to be compared [3]. Whereas previous

interleaving methods only measure the outcomes of items in the intersection of the original and interleaved lists, our proposed methods also measure the outcomes of items in the original lists but not in the interleaved list. We propose an interleaving method that selects items with equal probability for unbiased evaluation. With unequal selection probabilities, the evaluation might be biased due to confounding [8] between recommendation and potential outcomes, leading to inaccurate judgments of the recommendation models. We remove the possible bias by properly weighting the outcomes based on the inverse propensity score (IPS) method used in causal inference [16, 21]. This enables the use of a more general interleaving framework that only requires non-zero probabilities to be selected for any item in the original lists. As an instance of the framework, we propose a causal balanced interleaving method that balances the number of items chosen from the two compared lists. To verify the unbiasedness and efficiency of the proposed interleaving methods, we simulate online experiments to compare ranking models.

The contributions of this paper are summarized as follows.

- We propose the first interleaving methods to compare recommendation models in terms of their causal effect.
- We verify the unbiasedness and efficiency of the proposed methods through simulated online experiments.

2 RELATED WORK

2.1 Interleaving Methods

Interleaving is an online evaluation method for comparing two ranking models by observing user interactions with an interleaved list that is generated from lists ranked by the two models to be compared [3]. Several interleaving methods have been proposed for evaluating information retrieval systems. Balanced interleaving [14, 15] generates an interleaved list from two rankings to be compared such that the highest ranks in the interleaved list k_A and k_B from the two rankings A and B, respectively, are the same or different by at most one. Team draft interleaving [19] alternatively selects items from compared rankings, analogously to selecting teams for a friendly team-sports match. Probabilistic interleaving [9] selects items according to probabilities that depend on the item ranks. Optimized interleaving [18] makes the properties required for interleaving in information retrieval explicit and then generates interleaved lists by solving an optimization problem to fulfill those properties. Interleaving methods have been extended to multileaving that compare multiple rankings simultaneously [30, 31]. Multileaving has been also applied to the evaluation of a news recommender system [11]. The objective of previous interleaving methods is to evaluate how accurately the rankings reflect queries or user preferences, whereas our goal is to evaluate rankings in terms of the causal effect. To the best of our knowledge, at present there are no interleaving methods for causal effects.

2.2 Recommendation Methods for the Causal Effect

Recommendations can affect users' opinions [5] and induce users' actions [6, 13]. However, users' actions on recommended items could have occurred even without the recommendations [32]. Building recommendation models that target the causal effect is challenging because the ground truth data of causal effects are not observable [10]. One approach is to train prediction models for both recommended and non-recommended outcomes and then to rank the items based on the difference between the two predictions [1, 24]. Another approach is to optimize models directly for the causal effect. ULRMF and ULBPR [25] are respectively pointwise and pairwise optimization methods that use label transformations and training data samplings designed for causal effect optimization. DLCE [27] is an unbiased learning method for the causal effect that uses an IPS-based unbiased learning objective. There are also neighborhood methods for causal effects [28] that are based on a matching estimator in causal inference. These prior works on causal effects

evaluated methods offline and did not discuss protocols for online evaluation. In this study, we develop online evaluation methods and compare some of the aforementioned recommendation methods in simulated online experiments.

Another line of works in the area of causal recommendation aims for debiasing [4]. Several methods have been proposed to learn users' true preferences from biased (missing-not-at-random) feedback data [2, 23, 29, 33].[1] These methods can be regarded as predicting interactions with recommendations (i.e., Y_{ui}^{T}, defined in the next section). Hence, we can evaluate them using previous interleaving methods.

3 EVALUATION METHODS FOR THE CAUSAL EFFECT OF RECOMMENDATIONS

3.1 Causal Effect of Recommendations

In this subsection, we define the causal effect of recommendations. Let $\mathcal{U}$ and $\mathcal{I}$ be sets of users and items, respectively. Let $Y_{ui} \in \{0, 1\}$ denote the interaction (e.g., purchase or view) of user $u \in \mathcal{U}$ with item $i \in \mathcal{I}$. User interactions may differ depending on whether the item is recommended or not. We denote the binary indicator for the recommendation (also called the treatment assignment) by $Z_{ui} \in \{0, 1\}$. Let Y_{ui}^{T} and $Y_{ui}^{\mathrm{C}} \in \{0, 1\}$ be hypothetical user interactions (also called potential outcomes [22]) when item i is recommended to u ($Z_{ui} = 1$) and when it is not recommended ($Z_{ui} = 0$), respectively. The causal effect τ_{ui} of recommending item i to user u is defined as the difference between the two potential outcomes: $\tau_{ui} = Y_{ui}^{\mathrm{T}} - Y_{ui}^{\mathrm{C}}$, that takes ternary values, $\tau_{ui} \in \{-1, 0, 1\}$. Using potential outcomes, the observed interaction can be expressed as

$$Y_{ui} = Z_{ui} Y_{ui}^{\mathrm{T}} + (1 - Z_{ui}) Y_{ui}^{\mathrm{C}}. \tag{1}$$

$Y_{ui} = Y_{ui}^{\mathrm{T}}$ if i is recommended ($Z_{ui} = 1$) and $Y_{ui} = Y_{ui}^{\mathrm{C}}$ if it is not recommended ($Z_{ui} = 0$). Note that Y_{ui}^{T} or Y_{ui}^{C} cannot both be observed at a specific time; hence, τ_{ui} is not directly observable.

The recommendation model A generates a recommendation list L_u^A for each user. The average causal effect of model A is then defined as

$$\tau_A = \mathbb{E}[\tau_{ui} | i \in L_u^A, u \in \mathcal{U}]. \tag{2}$$

In this work, we evaluate models using the above metric.[2] That is, when comparing two models, we regard A to superior to B when $\tau_A > \tau_B$.

3.2 A/B testing for the Causal Effect

For A/B testing, we randomly select non-overlapping subsets of users $\mathcal{S}_A$ and $\mathcal{S}_B$ (i.e., $\mathcal{S}_A, \mathcal{S}_B \subset \mathcal{U}$ and $\mathcal{S}_A \cap \mathcal{S}_B = \emptyset$) and apply models A and B to each subset. Let $n = |L_u^A| = |L_u^B|$ be the size of the recommendation list, which we assume to be constant. The subset average causal effect is then defined as

$$\hat{\tau}_A = \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \sum_{i \in L_u^A} \tau_{ui}. \tag{3}$$

This converges to τ_A as $|\mathcal{S}_A|$ increases.

The typical evaluation metrics for A/B testing are either based on total user interactions (such as sales or user engagement) or only on interactions with recommended lists (such as click-through rates or conversion rates) [13].

---

[1]Note that CausE proposed by Bonner and Vasile [2] can be used for the causal effect ranking [25], although their original work tackles unbiased prediction of Y_{ui}^{T} and only refers to the prediction of τ_{ui}. Wang et al. [33] suggested that they want to recommend items that have low probability of exposure and that would be rated high if exposed. Their approach might be regarded as indirectly targeting the causal effect of recommendations, assuming that recommendations increase exposures. To take this approach, it might be also important to model the influence of recommendations on exposures [26].

[2]This metric is identical to the causal precision@n in [27].

Here we show that the former is a valid evaluation for the causal effect. The total user interactions divided by the number of recommendations can be expressed as

$$
\begin{aligned}
\hat{Y}_A^{\text{total}} &= \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \sum_{i \in \mathcal{I}} Y_{ui} \\
&= \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \left(\sum_{i \in L_u^A} Y_{ui}^{\text{T}} + \sum_{i \in \mathcal{I} \setminus L_u^A} Y_{ui}^{\text{C}} \right) = \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \left(\sum_{i \in L_u^A} (\tau_{ui} + Y_{ui}^{\text{C}}) + \sum_{i \in \mathcal{I} \setminus L_u^A} Y_{ui}^{\text{C}} \right) \\
&= \hat{\tau}_A + \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \sum_{i \in \mathcal{I}} Y_{ui}^{\text{C}}.
\end{aligned}
\tag{4}
$$

Because the rightmost term in the final equation does not depend on the model, we can compare $\hat{\tau}_A$ and $\hat{\tau}_B$ by comparing $\hat{Y}_A^{\text{total}}$ and $\hat{Y}_B^{\text{total}}$. On the other hand, the average interactions with the recommended lists can be expressed as

$$
\begin{aligned}
\hat{Y}_A^{\text{list}} &= \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \sum_{i \in L_u^A} Y_{ui} \\
&= \frac{1}{n|\mathcal{S}_A|} \sum_{u \in \mathcal{S}_A} \sum_{i \in L_u^A} Y_{ui}^{\text{T}} \neq \hat{\tau}_A.
\end{aligned}
\tag{5}
$$

Hence, the evaluation based only on interactions with recommended lists is not valid testing for the causal effect.

Although A/B testing with Eq. (4) can be used for unbiased model comparisons, it may have large variance due to the variability in natural user behaviors (i.e., the potential outcomes under no recommendations, Y_{ui}^{C}). If users in $\mathcal{S}_A$ tend to purchase more items than those in $\mathcal{S}_B$, $\sum_{u \in \mathcal{S}_A} \sum_{i \in \mathcal{I}} Y_{ui}^{\text{C}}$ becomes larger than $\sum_{u \in \mathcal{S}_B} \sum_{i \in \mathcal{I}} Y_{ui}^{\text{C}}$, thereby altering the comparison in Eq. (4). To minimize such discrepancies, a sufficiently large number of users need to be recruited for A/B testing. We thus introduce more efficient evaluation methods in the next subsection.

3.3 Interleaving for the Causal Effect

In this subsection, we propose interleaving methods for the online evaluation of the causal effects of recommendations. Previous interleaving methods only measure outcomes in the interleaved lists: they only include Y_{ui}^{T} and lack information on Y_{ui}^{C}. Further, if the item selection for the interleaved list is not randomized controlled, the naive estimate from the observed outcomes might be biased due to the confounding between recommendations and potential outcomes. We need to remedy the bias for valid comparison.

Here we describe the problem setting of interleaving for the causal effect. For each user u, we construct the interleaved list L_u from the compared lists L_u^A and L_u^B. We observe outcomes $\{Y_{ui}\}$ for all items $i \in \mathcal{I}$. Note that $Y_{ui} = Y_{ui}^{\text{T}}$ if item i is in the interleaved list ($i \in L_u$ or equivalently, $Z_{ui} = 1$) and $Y_{ui} = Y_{ui}^{\text{C}}$ if it is not in the list ($i \in \mathcal{I} \setminus L_u$ or equivalently, $Z_{ui} = 0$). We want to compare the average causal effects of lists L_u^A and L_u^B:

$$
\tau_{L_u^A} = \frac{1}{n} \sum_{i \in L_u^A} \tau_{ui}, \quad \tau_{L_u^B} = \frac{1}{n} \sum_{i \in L_u^B} \tau_{ui}.
\tag{6}
$$

We need to estimate the above values from observed outcomes because we cannot directly observe τ_{ui}.

If the items in L_u^A and L_u^B are randomly assigned to the interleaved list independent of the potential outcomes, that is, $\left(Y_{ui}^{\text{T}}, Y_{ui}^{\text{C}} \right) \perp Z_{ui}$, the case can be regarded as a randomized controlled trial (RCT) [12, 22].[3] We can then simply estimate

---

[3]For our interleaving methods, the independence is required only for the items in the union of L_u^A and L_u^B.

$\tau_{L_u^A}$ as the difference in average outcomes for items on and not on the interleaved list:

$$\left(\hat{\tau}_{L_u^A}\right)_{\text{RCT}} = \frac{1}{|L_u^A \cap L_u|} \sum_{i \in L_u^A \cap L_u} Y_{ui} - \frac{1}{|L_u^A \setminus L_u|} \sum_{i \in L_u^A \setminus L_u} Y_{ui}. \tag{7}$$

One way to realize such a randomized assignment is to select n items from $L_u^A \cup L_u^B$ with equal probability: $p = n/|L_u^A \cup L_u^B|$. We call this method equal probability interleaving (EPI).

The independence requirement heavily restricts the potential design space of interleaving methods. We thus derive estimates that are applicable to more general cases. Denote the probability (also called the propensity) of being included in the interleaved list L_u by $p_{ui} = \mathbb{E}[Z_{ui} = 1|X_{ui}]$. We assume that 1) the covariates X_{ui} contain all confounders of $\left(Y_{ui}^{\text{T}}, Y_{ui}^{\text{C}}\right)$ and Z_{ui}, and 2) the treatment assignment is not deterministic ($0 < p_{ui} < 1$ for $i \in L_u^A \cup L_u^B$).[4] Assumption 1 is equivalent to conditional independence: $\left(Y_{ui}^{\text{T}}, Y_{ui}^{\text{C}}\right) \perp Z_{ui}|X_{ui}$. When we design an interleaving method, we know the covariates that affect Z_{ui} and Assumption 1 can always be satisfied.[5] Therefore, the only restriction for interleaving methods is Assumption 2 (also called *positivity*).

Under these assumptions, we can construct an unbiased estimator using IPS weighting [16]:

$$\left(\hat{\tau}_{L_u^A}\right)_{\text{IPS}} = \frac{1}{n} \sum_{i \in L_u^A \cap L_u} \frac{Y_{ui}}{p_{ui}} - \frac{1}{n} \sum_{i \in L_u^A \setminus L_u} \frac{Y_{ui}}{1 - p_{ui}} = \frac{1}{n} \sum_{i \in L_u^A} \left(\frac{Z_{ui} Y_{ui}}{p_{ui}} - \frac{(1 - Z_{ui}) Y_{ui}}{1 - p_{ui}} \right). \tag{8}$$

This estimator is unbiased since

$$\mathbb{E}\left[\frac{Z_{ui} Y_{ui}}{p_{ui}} - \frac{(1 - Z_{ui}) Y_{ui}}{1 - p_{ui}} \middle| X_{ui} \right] = \mathbb{E}\left[\frac{Z_{ui} Y_{ui}^{\text{T}}}{p_{ui}} - \frac{(1 - Z_{ui}) Y_{ui}^{\text{C}}}{1 - p_{ui}} \middle| X_{ui} \right] = \frac{\mathbb{E}[Z_{ui}|X_{ui}] Y_{ui}^{\text{T}}}{p_{ui}} - \frac{\mathbb{E}[(1 - Z_{ui})|X_{ui}] Y_{ui}^{\text{C}}}{1 - p_{ui}}$$

$$= \frac{p_{ui} Y_{ui}^{\text{T}}}{p_{ui}} - \frac{(1 - p_{ui}) Y_{ui}^{\text{C}}}{1 - p_{ui}} = \tau_{ui}. \tag{9}$$

We propose a general framework for interleaving as follows.

(1) Construct interleaved lists $\{L_u\}$ using an interleaving method that satisfies positivity (Assumption 2).
(2) Conduct online experiments and obtain outcomes $\{Y_{ui}\}$.
(3) Estimate $\tau_{L_u^A}$ and $\tau_{L_u^B}$ by Eq. (8) and compare them.

As an example of a valid interleaving method that satisfies positivity, we propose causal balanced interleaving (CBI), the pseudo-code for which is shown in Algorithm 1. CBI alternatively selects items from each list to balance the items chosen from each list. The item choice in each round is not deterministic in order to satisfy the positivity required for causal effect estimates. The propensity depends on whether an item is in the intersection, $\mathbf{1}(i \in L_u^A \cap L_u^B)$. If an item is included in both lists, it has a greater probability of being chosen. The propensity also depends on the cardinality of the union of the compared lists, $|L_u^A \cup L_u^B|$, because smaller cardinality implies that each item has a greater chance of being selected. The possible values of the covariates are limited: $\mathbf{1}(i \in L_u^A \cap L_u^B)$ is binary and $|L_u^A \cup L_u^B| \in [n, 2n]$. Hence, we can easily compute the propensity numerically by repeating Algorithm 1 a sufficient number of times and recording Z_{ui} for each combination of covariates.

---

[4]Taken together, these two assumptions are called *strongly ignorable treatment assignment* [21].
[5]Confounders are covariates that affect both $\left(Y_{ui}^{\text{T}}, Y_{ui}^{\text{C}}\right)$ and Z_{ui}, and they are subsets of covariates that affect Z_{ui}. Hence, including the latter in X_{ui} is a sufficient condition for Assumption 1.

Algorithm 1: Causal Balanced Interleaving (*CBI*).

Input: Compared lists L_u^A and L_u^B
Output: Interleaved list L_u with size n ($n = |L_u^A| = |L_u^B|$)

```
1  L_u ← ()                                              // initialize interleaved list
2  ANext ← RandBit() = 1                     // Randomly select whether starting from A or B
3  while |L_u| < n do
4  │   if ANext then
5  │   │   L_u ← L_u + RandomChoiceFrom(L_u^A\L_u)  // Randomly choose one item from L_u^A not yet in L_u
6  │   │   ANext ← False
7  │   else
8  │   │   L_u ← L_u + RandomChoiceFrom(L_u^B\L_u)  // Randomly choose one item from L_u^B not yet in L_u
9  │   │   ANext ← True
10 return L_u
```

---

4 EXPERIMENTS

4.1 Experimental Setup

We experimented with the following online evaluation methods.[6]

- AB-total: A/B testing evaluated by the total user interactions, as expressed in Eq. (4).
- AB-list: A/B testing evaluated by user interactions only with items on the recommended list, as in Eq. (5).
- EPI-RCT: Interleaving to select items from $L_u^A \cup L_u^B$ with equal probability and evaluation using Eq. (7).
- CBI-RCT: Interleaving by Algorithm 1 and evaluation using Eq. (7), that is, no bias correction by IPS.
- CBI-IPS: Interleaving by Algorithm 1 and evaluation using Eq. (8).

Through the experiments, we aim to answer the following research questions: RQ1) Which method produces valid (unbiased) estimates of the true differences in average causal effects (4.2.1)?, and RQ2) Are the proposed interleaving methods more efficient (do they require fewer experimental users) than AB testing (4.2.2)? We first prepared semi-synthetic datasets that contain both potential outcomes Y_{ui}^T and Y_{ui}^C for all user-item pairs. Because we observe $Y_{ui} = Y_{ui}^T$ if $Z_{ui} = 1$ and $Y_{ui} = Y_{ui}^C$ if $Z_{ui} = 0$, both potential outcomes are necessary to simulate user outcomes under various ranking models and online evaluation methods. Following the procedure described in [28], we generated two datasets: one is based on the Dunnhumby dataset,[7] and the other is based on the MovieLens-1M (ML-1M) dataset [7].[8] The detail and rationale of ML one are described in Section 5.1 of [28] and that of DH one are described in 5.1.1 of [27]. Each dataset is comprised of independently generated training and testing data. The testing data were used to simulate online evaluation, and the training data were used to train the following models:[9] the causality-aware user-based neighborhood methods (CUBN) with outcome similarity (-O) and treatment similarity (-T) [28], the uplift-based pointwise and pairwise learning methods (ULRMF and ULBPR) [25], the Bayesian personalized ranking method (BPR) [20], and the user-based neighborhood method (UBN) [17]. We compared two models among CUBN-T, ULRMF, BPR on the Dunnhumby data and two models among CUBN-O, ULBPR, UBN on the ML-1M data.[10] The average causal effect $\overline{\tau_{L_u^{model}}}$ and the average

---

[6]For reproducibility, the code is available at https://github.com/masatoh73/causal-interleaving.
[7]https://www.dunnhumby.com/careers/engineering/sourcefiles
[8]https://grouplens.org/datasets/movielens
[9]We used hyper-parameters for CP@10, described in the ancillary files at http://arxiv.org/abs/2012.09442.
[10]We intended to compare models of different families, i.e., one of {CUBN-T, CUBN-O} with one of {ULBPR, ULRMF}.

treated outcomes $\overline{Y^{\mathrm{T}}_{L^{model}_u}}$ of the trained models are listed in Table 1. The superior models in terms of the average causal effect do not necessarily have higher average treated outcomes. That is, we may mistakenly select a poor model in terms of the causal effect if we only evaluate the outcomes of the recommended items.

Table 1. Averages of causal effect and potential outcomes under treatment with recommendation lists of size $n = 10$.

| | Dunnhumby-Original | | | MovieLens-1M | | |
	CUBN-T	ULRMF	BPR	CUBN-O	ULBPR	UBN
$\overline{\tau_{L^{model}_u}}$	0.0507	0.0347	0.0295	0.332	0.280	-0.186
$\overline{Y^{\mathrm{T}}_{L^{model}_u}}$	0.1359	0.1396	0.1869	0.341	0.285	0.308

Our protocol for simulating online experiments is the following. First, we randomly select a subset of users and generate lists L^A_u, L^B_u using compared models. For the A/B testing methods (AB-total, AB-list), we further split the subset into two groups: $\mathcal{S}_A$ and $\mathcal{S}_B$, and $\{L^A_u\}$ and $\{L^B_u\}$ are recommended for each group, respectively. For the interleaving methods (EPI-RCT, CBI-RCT, CBI-IPS), we generate interleaved recommendation lists using EPI or CBI. In the simulation, *recommendation* means that Z_{ui} is set to 1, and user outcomes $\{Y_{ui}\}$ are *observed* by calculating $Y_{ui} = Z_{ui}Y^{\mathrm{T}}_{ui} + (1-Z_{ui})Y^{\mathrm{C}}_{ui}$ with potential outcomes Y^{T}_{ui} and Y^{C}_{ui}. Using the observed outcomes, we estimate the difference in the average causal effects of the compared models: $\tau_A - \tau_B$. We repeated the above protocol 10,000 times and recorded the estimated differences using each online evaluation method. The size of recommendation list was set to 10.

4.2 Results and Discussion

4.2.1 Validity of the evaluation methods. We evaluated the validity of the online evaluation methods using random subsets of 1,000 users. The means and standard deviations of the estimated differences are shown in Table 2. The means obtained by EPI-RCT and CBI-IPS are close to the true differences. The means obtained by AB-total are also close to the true value for Dunnhumby but deviate slightly for ML-1M. The AB-list often yields estimates that differ substantially from the true values but are similar to the differences in treated outcomes, $\overline{Y^{\mathrm{T}}_{L^{model}_u}}$, as shown in Table 1. This is expected because the AB-list evaluates Y^{T}_{ui}, not τ_{ui}, as expressed in Eq. (5). Further, the CBI-RCT estimates also deviate from the true differences in most cases.[11] This is due to the bias induced by the uneven probability of recommendation in interleaving. Conversely, CBI-IPS successfully removes the bias and produces estimates centered around the true values.

Table 2. Estimated differences between the causal effects of the compared models (mean ± standard deviations for 10,000 simulated runs). The results highlighted in bold indicate that the true values are within the 95% confidence intervals of the mean estimates.

| | Dunnhumby-Original | | | MovieLens-1M | | |
	CUBN-T & BPR	CUBN-T & ULRMF	ULRMF & BPR	CUBN-O & UBN	CUBN-O & ULBPR	ULBPR & UBN
Truth	0.0212	0.0160	0.0052	0.5177	0.0512	0.4665
AB-total	**0.0210** ± 0.0399	**0.0159** ± 0.0399	**0.0051** ± 0.0397	**0.5301** ± 1.2048	**0.0635** ± 1.2102	**0.4789** ± 1.2052
AB-list	-0.0510 ± 0.0071	-0.0037 ± 0.0065	-0.0471 ± 0.0073	0.0325 ± 0.0104	0.0550 ± 0.0104	-0.0226 ± 0.0100
EPI-RCT	**0.0212** ± 0.0069	**0.0159** ± 0.0075	**0.0053** ± 0.0076	**0.5178** ± 0.0137	**0.0512** ± 0.0083	**0.4666** ± 0.0135
CBI-RCT	0.0429 ± 0.0067	0.0192 ± 0.0067	0.0188 ± 0.0076	**0.5179** ± 0.0126	0.0444 ± 0.0066	**0.4667** ± 0.0126
CBI-IPS	**0.0213** ± 0.0063	**0.0160** ± 0.0066	**0.0051** ± 0.0070	**0.5179** ± 0.0126	**0.0512** ± 0.0075	**0.4667** ± 0.0126

---

[11]In the comparisons of CUBN-O & UBN and ULBPR & UBN, the results of CBI-RCT and CBI-IPS are identical. There was no overlaps between L^A_u and L^B_u in these comparisons, and the propensity was constant. Hence, IPS was not necessary, and CBI-RCT and CBI-IPS were equivalent.

4.2.2 Efficiency of the interleaving methods. We compared the efficiency of AB-total, EPI-RCT, and CBI-IPS, all of which were shown to be valid in the previous section. We simulated user subsets of various sizes in {10, 14, 20, 30, 50, 70, 100, 140, 200, 300, 500, 700, 1000, 1400, 2000} and evaluated the ratio of false judgments (when the sign of the estimated difference is the opposite of the truth). Figure 1 shows the ratio of false judgments according to the number of users. As the number of users increases, the false ratios of CBI-IPS and EPI-RCT decrease more rapidly than that of AB-total does. For the Dunnhumby dataset, AB-total requires around 30 times more users than CBI-IPS and EPI-RCT to achieve the same false ratio. For the ML-1M dataset, AB-total did not reach the same false ratio in the experimental range of subset sizes. These results demonstrate the superior efficiency of the proposed interleaving methods. Furthermore, CBI-IPS tends to be slightly more efficient than EPI-RCT, as expected from the smaller standard deviations shown in Table 2. This is probably because the number of items selected from the compared lists is balanced in this interleaving method.
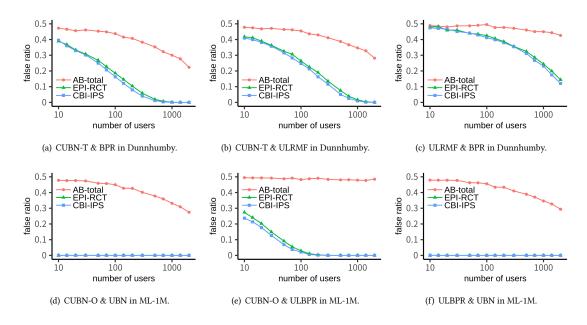

(a) CUBN-T & BPR in Dunnhumby.   (b) CUBN-T & ULRMF in Dunnhumby.   (c) ULRMF & BPR in Dunnhumby.

(d) CUBN-O & UBN in ML-1M.   (e) CUBN-O & ULBPR in ML-1M.   (f) ULBPR & UBN in ML-1M.

Fig. 1. Dependence on the number of users.

5 CONCLUSIONS

In this paper, we proposed the first interleaving methods for comparing recommender models in terms of causal effects. To realize unbiased model comparisons, our methods either select items with equal probabilities or weight the outcomes using IPS. We simulated online experiments and verified that our interleaving methods and an A/B testing method are unbiased and that our interleaving methods are largely more efficient than the A/B testing method. In the future, we plan to extend our methods to multileaving. Online experimentation in real recommendation services will also be important for future work.

REFERENCES

[1] Anand V Bodapati. 2008. Recommendation systems with purchase data. *Journal of marketing research* 45, 1 (2008), 77–93.

[2] Stephen Bonner and Flavian Vasile. 2018. Causal Embeddings for Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 104–112. https://doi.org/10.1145/3240323.3240360

[3] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-Scale Validation and Analysis of Interleaved Search Evaluation. *ACM Trans. Inf. Syst.* 30, 1, Article 6 (March 2012), 41 pages. https://doi.org/10.1145/2094072.2094078

[4] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240* (2020).

[5] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is Seeing Believing? How Recommender System Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 585–592. https://doi.org/10.1145/642611.642713

[6] M. Benjamin Dias, Dominique Locher, Ming Li, Wael El-Deredy, and Paulo J.G. Lisboa. 2008. The Value of Personalised Recommender Systems to E-Business: A Case Study. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (Lausanne, Switzerland) *(RecSys '08)*. Association for Computing Machinery, New York, NY, USA, 291–294. https://doi.org/10.1145/1454008.1454054

[7] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[8] MA Hernán and JM Robins. 2020. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC* (2020).

[9] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. 2011. A Probabilistic Method for Inferring Preferences from Clicks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) *(CIKM '11)*. Association for Computing Machinery, New York, NY, USA, 249–258. https://doi.org/10.1145/2063576.2063618

[10] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

[11] Kojiro Iizuka, Takeshi Yoneda, and Yoshifumi Seki. 2019. Greedy Optimized Multileaving for Personalization. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 413–417. https://doi.org/10.1145/3298689.3347008

[12] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York, NY, USA.

[13] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.* 10, 4, Article 16 (Dec. 2019), 23 pages. https://doi.org/10.1145/3370082

[14] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada) *(KDD '02)*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/775047.775067

[15] Thorsten Joachims. 2003. Evaluating Retrieval Performance Using Clickthrough Data. In *Text Mining, Theoretical Aspects and Applications*, Jürgen Franke, Gholamreza Nakhaeizadeh, and Ingrid Renz (Eds.). Physica-Verlag, 79–96.

[16] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[17] Xia Ning, Christian Desrosiers, and George Karypis. 2015. *A Comprehensive Survey of Neighborhood-Based Recommendation Methods.* Springer US, Boston, MA, 37–76.

[18] Filip Radlinski and Nick Craswell. 2013. Optimized Interleaving for Online Retrieval Evaluation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (Rome, Italy) *(WSDM '13)*. Association for Computing Machinery, New York, NY, USA, 245–254. https://doi.org/10.1145/2433396.2433429

[19] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA) *(CIKM '08)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/1458082.1458092

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

[21] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[22] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.

[23] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 501–509. https://doi.org/10.1145/3336191.3371783

[24] Masahiro Sato, Hidetaka Izumo, and Takashi Sonoda. 2016. Modeling Individual Users' Responsiveness to Maximize Recommendation Impact. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (Halifax, Nova Scotia, Canada) *(UMAP '16)*. ACM, New York, NY, USA, 259–267. https://doi.org/10.1145/2930238.2930259

[25] Masahiro Sato, Janmajay Singh, Sho Takemori, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2019. Uplift-Based Evaluation and Optimization of Recommenders. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 296–304. https://doi.org/10.1145/3298689.3347018

[26] Masahiro Sato, Janmajay Singh, Sho Takemori, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2020. Modeling User Exposure with Recommendation Influence. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (Brno, Czech Republic) *(SAC '20)*. Association for Computing Machinery, New York, NY, USA, 1461–1464. https://doi.org/10.1145/3341105.3375760

[27] Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. 2020. Unbiased Learning for the Causal Effect of Recommendation. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 378–387. https://doi.org/10.1145/3383313.3412261

[28] Masahiro Sato, Sho Takemori, Janmajay Singh, and Qian Zhang. 2021. Causality-Aware Neighborhood Methods for Recommender Systems. (2021), 603–618. https://doi.org/10.1007/978-3-030-72113-8_40

[29] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1670–1679.

[30] Anne Schuth, Robert-Jan Bruintjes, Fritjof Buüttner, Joost van Doorn, Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling, Jos van der Velde, Roger Wechsler, David Woudenberg, and Maarten de Rijke. 2015. Probabilistic Multileave for Online Retrieval Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 955–958. https://doi.org/10.1145/2766462.2767838

[31] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. 2014. Multileaved Comparisons for Fast Online Evaluation. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. Association for Computing Machinery, New York, NY, USA, 71–80. https://doi.org/10.1145/2661829.2661952

[32] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (Portland, Oregon, USA) *(EC '15)*. ACM, New York, NY, USA, 453–470. https://doi.org/10.1145/2764468.2764488

[33] Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. 2020. Causal Inference for Recommender Systems. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 426–431. https://doi.org/10.1145/3383313.3412225