Predicting Search Satisfaction Metrics with Interleaved Comparisons

Anne Schuth*
University of Amsterdam
anne.schuth@uva.nl

Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

Filip Radlinski
Microsoft
filiprad@microsoft.com

ABSTRACT

The gold standard for online retrieval evaluation is AB testing. Rooted in the idea of a controlled experiment, AB tests compare the performance of an experimental system (treatment) on one sample of the user population, to that of a baseline system (control) on another sample. Given an online evaluation metric that accurately reflects user satisfaction, these tests enjoy high validity. However, due to the high variance across users, these comparisons often have low sensitivity, requiring millions of queries to detect statistically significant differences between systems. Interleaving is an alternative online evaluation approach, where each user is presented with a combination of results from both the control and treatment systems. Compared to AB tests, interleaving has been shown to be substantially more sensitive. However, interleaving methods have so far focused on user clicks only, and lack support for more sophisticated user satisfaction metrics as used in AB testing.

In this paper we present the first method for integrating user satisfaction metrics with interleaving. We show how interleaving can be extended to (1) directly match user signals and parameters of AB metrics, and (2) how parameterized interleaving credit functions can be automatically calibrated to predict AB outcomes. We also develop a new method for estimating the relative sensitivity of interleaving and AB metrics, and show that our interleaving credit functions improve agreement with AB metrics without sacrificing sensitivity. Our results, using 38 large-scale online experiments encompassing over 3 billion clicks in a web search setting, demonstrate up to a 22% improvement in agreement with AB metrics (constituting over a 50% error reduction), while maintaining sensitivity of one to two orders of magnitude above the AB tests. This paves the way towards more sensitive and accurate online evaluation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords

Information retrieval; Evaluation; Interleaved comparisons

*Most of this work done during internship at Microsoft Research.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767695.

1. INTRODUCTION

Evaluation has long played a key role in information retrieval (IR). Most commonly, systems are evaluated following the Cranfield approach [25]. Using this approach, systems are evaluated in terms of document relevance for given queries, which is assessed by trained experts. While the Cranfield paradigm ensures high internal validity and repeatability of experiments, it has been shown that the users' search success and satisfaction with an IR system are not always accurately reflected by standard IR metrics [29, 31]. One reason is that the relevance judges typically do not assess queries and documents that reflect their own information needs, and have to make assumptions about relevance from an assumed users point of view. Because the true information need can be difficult to assess, this can cause substantial biases [10, 30, 36].

To address the gap between offline evaluation and true use of IR systems, *online* evaluation has been used to directly measure observable user behavior on alternative systems. The biggest challenge for online evaluation is to identify metrics that accurately reflect user satisfaction. This has motivated a large amount of research on online metrics. While early online evaluation focused on simple metrics such as click-through rate (CTR, the fraction of queries for which users click a result) or the ranks of clicked documents [14], more sophisticated metrics have been recently developed. These include observing which results users skip over [34], the time between search engine visits [3], and focusing on "satisfied" (long-duration) clicks (which we refer to as SAT clicks) [4].

Given an IR system and an appropriate online metric, the standard experimental procedure for comparing systems is AB testing [17]. This means that a controlled experiment is conducted on users of a running system. A random sample of users is exposed to the treatment system, a second sample is exposed to the control system. Given that the assignment to systems is random and the experimental units (e.g., users) can be assumed independent of each other, any differences in online performance measured on the two samples can be attributed to differences between treatment and control system. If the measured differences are statistically significant, we can make highly confident decisions on which system to deploy. Unfortunately, the variance in user behavior is typically high, which results in low sensitivity of such AB comparisons. This means that, to reach high confidence levels, large samples need to be collected over a long period of time (e.g., millions of samples) [1]. Considering the effects of exposing users to potentially lower quality systems over long periods of time, it can be seen that AB comparisons can be extremely expensive.

An alternative online evaluation approach, interleaving, was developed to improve on the AB design [14]. It avoids many of the sources of variance by combining results from both the treatment and control systems, for all queries. In particular, the results returned

by the two systems are combined in a way that is fair to both, in the sense that neither system would be preferred in expectation if users were to click on documents at random. Observed user clicks on the combined result list are then credited to one of the systems to infer which system would be preferred by the user [24]. In comparison to AB tests, interleaved comparisons have been shown to be substantially more sensitive. For instance, in an empirical comparison of five AB tests and corresponding interleaving experiments, Chapelle et al. [1] observed that AB experiments required 145 times more data than interleaving to achieve statistical significance.

While high sensitivity makes interleaving very attractive for online evaluation, existing methods have primarily focused on observed clicks, and ignore the richer user satisfaction signals that have been incorporated into AB metrics. As a result, it was unclear to what degree interleaved comparisons agree with user satisfaction, as measured by specifically designed AB tests. This is challenging to address because by their very nature interleaving methods change rankings and attribute credit in a non straight forward way, making it far from trivial to align them with AB metrics. This paper is the first to address this limitation of interleaving methods.

In particular, we make the following contributions:

Sensitivity: Starting with existing AB and interleaving metrics (defined in Section 3), we propose a new, statistical method for assessing the sensitivity of these metrics from estimated effect sizes (Section 4). The resulting method allows a detailed comparison between metrics in terms of the power of statistical tests at varying sample sizes. Our analysis shows that AB tests typically require two orders of magnitude more data than interleaved comparisons.

Agreement: Turning to the agreement between existing metrics, we find that current interleaved comparisons achieve from random up to 76% agreement with AB user satisfaction metrics.

Credit Formulation: Motivated by the results of our analysis, we propose novel interleaving *credit functions* that are (a) designed to closely match the implementation and parameters of AB metrics, or (b) are parameterized to allow optimization towards agreement with arbitrary AB metrics (Section 5). We further propose the first approach for automatically maximizing agreement between such parameterized interleaving credit functions and AB metrics.

Optimization: We demonstrate that interleaving credit functions can be automatically optimized, and that learned parameters generalize to unseen experiments. These results demonstrate for the first time that interleaving can be augmented with user satisfaction metrics, to accurately predict the outcomes of AB comparisons that would require one to two orders of magnitude more data.

Large Scale Evaluation: Finally, our empirical results, obtained from experiments with 3 billion user impressions and 38 paired (AB and interleaving) experiments demonstrate the effectiveness of our proposed approach (Section 6). In particular, we achieve agreement of up to 87%, while maintaining high sensitivity.

2. RELATED WORK

In this section, we discuss prior work on measuring user satisfaction with online IR systems, starting with absolute relevance metrics: Metrics that measure a single number for a given ranking system (2.1). We then present existing results on online evaluation using interleaving (2.2). Finally we discuss approaches for optimizing online evaluation metrics as relevant to this paper (2.3).

2.1 Absolute User Metrics

The most common type of online evaluation today is AB testing, where online performance is estimated on two independent samples of users: One exposed to a treatment system, the other exposed to a control system [17]. Standard assumptions allow experimenters to obtain unbiased online performance estimates, and confidence estimates or hypothesis testing are available via statistical methods such as the two-sample t-test. For instance, this methodology has been effectively used to compare systems in terms of CTR (e.g., for news recommendation [20]).

While simple to measure, CTR has been shown to be a poor metric for measuring user satisfaction in search [15]. Consequently, a large body of work has developed online metrics that more accurately measure search satisfaction. An established signal is dwell time, where clicks followed by only short visits to the corresponding result document are considered "unsatisfied", i.e., the user is unlikely to have found the document as relevant [36]. Moving beyond a single time threshold for identifying user satisfaction, sophisticated click satisfaction classifiers combine a range of user signals, and have been shown to accurately detect satisfied clicks [16].

Other proposed online metrics also consider the effect of tabbed browsing (opening several results in browser tabs in quick succession) [13]. Conversely, the lack of a click (abandonment) is often taken as a signal of a lack of relevance, but this interpretation has posed a challenge for evaluating richer search engine result pages, where relevant information may be presented directly, without the need to click. A number of papers have proposed methods to determine when abandonment indicates satisfaction [20, 28]. Follow-on queries can also be considered indicative of a lack of success [8], as can skipping results be indicative of incorrect result order [34]. Of course, to accurately interpret user clicks, we must also consider which results users examined. If a user never looked at a search result, their lack of engagement on this result cannot be indicative of low relevance. A number of studies have shown that mouse movement can be an indicator of user examination of search results, and of specific sections within search results [2, 5, 6, 9, 35]. Similarly, in a mobile setting recording how long each part of the screen is visible can be considered an indicator of relevance [18].

Finally, although the majority of online evaluation has focused on user satisfaction for individual queries, it has been argued that the correct unit of measurement is the user session, or a search task. A number of session based metrics have been proposed [7, 33].

In this paper we specifically focus on AB metrics that capture click-level search satisfaction [16] (cf., Section 3.1).

2.2 Interleaved Comparisons

While providing flexibility and control, AB comparisons typically require a large number of observations. Given typical differences in IR system performance in state of the art systems, many AB metrics have been found to require millions of users [1]. Interleaved methods, originally proposed by Joachims [14], reduce the variance of measurement by combining documents retrieved by both the control and the treatment system. Projecting user clicks on the resulting interleaved document lists back to the original document rankings is then taken as an estimate of which system would be preferred by the user. This mixing substantially reduces variance and was found to reduce required sample sizes by up to two orders of magnitude [1, 24]. A variety of interleaving approaches have been proposed (e.g. [11, 14, 23, 24, 26, 27]). The most frequently used interleaving algorithm is Team Draft Interleaving (TDI) [24]. We use TDI as our baseline. It is described in detail in Section 3.2.

2.3 Optimizing Interleaving Metrics

The interleaving approaches described above measure which ranker is more likely to attract user clicks in a fair, paired comparison. However, as described in Section 2.1, raw clicks can be misleading. Previous research has shown that with interleaving there may be biases due to highlighting in search result titles [38] and other caption effects such as title and snippet length [12]. Proposed methods to mediate these biases were shown to improve agreement

Table 1: AB metrics implemented as ground truth for comparisons with interleaving. See Section 3.1 for notation.

AB Metric	Description	Implementation $\frac{1}{Q_A}\sum_{q\in Q_A}\ldots$
AB	Number of queries that received at least one click.	$\mathbf{1}(\lvert C^q\rvert > 1)$
AB@1	Number of queries that received at least one click on the first position.	$\mathbf{1}((\sum_{c\in C^q}\mathbf{1}(\text{rank}(c)=1)) > 1)$
AB_S	Number of queries that received at least one SAT click.	$\mathbf{1}((\sum_{c\in C^q}\mathbf{1}(\text{is\_sat}(c))) > 1)$
AB_S@1	Number of queries that received at least one SAT click on the first position.	$\mathbf{1}((\sum_{c\in C^q}\mathbf{1}(\text{rank}(c)=1)\cdot\mathbf{1}(\text{is\_sat}(c))) > 1)$
AB_T	Time from the query being issued until the first click.	$\min_{c\in C^q}\text{time}(c)$
AB_T@1	Time to the first click on the top position.	$\min_{c\in C^q}\text{time}(c)\cdot\mathbf{1}(\text{rank}(c)=1)$
$AB_{T,S}$	Time to the first click classified as SAT.	$\min_{c\in C^q}\text{time}(c)\cdot\mathbf{1}(\text{is\_sat}(c))$
$AB_{T,S}$@1	Time to the first click on the top position classified as SAT.	$\min_{c\in C^q}\text{time}(c)\cdot\mathbf{1}(\text{rank}(c)=1)\cdot\mathbf{1}(\text{is\_sat}(c))$

with offline evaluation [12, 38], but optimizing agreement with online metrics remains an open challenge.

Also, the above approaches may improve interleaving by removing some click bias, they still aim to be *unbiased* rather than *predictive of satisfaction*. In this paper, we show how to create an interleaving evaluation that instead aims to predict the outcome of an AB experiment for any given AB metric, while maintaining the sensitivity improvements of interleaving. In particular, we take into account whether clicks are indicators of success by reimplementing the classifier learned by Kim et al. [16].

The goal of this paper is also related to prior work on optimizing the sensitivity of interleaving algorithms, where interleaving algorithms were learned to be more statistically powerful [37], or to satisfy given choices about the value of any given preference observed [23]. In contrast, our work is the first that focuses on optimizing "correctness" of an interleaving outcome as captured in terms of agreement with AB metrics, while maintaining high sensitivity. Our results show that in this way agreement between interleaving and any given AB metric can be dramatically improved.

3. BACKGROUND

In this section we describe the most commonly used AB metrics, and the interleaved evaluation approach that we build on in the remainder of this paper. We will take the presented AB metrics as the *ground truth* user satisfaction metrics we aim to predict with much smaller interleaving samples.

3.1 Common AB Metrics

As described above, a large number of AB metrics have been developed. Most have in common that clicks are the basic observed interaction with users, thus this is our focus too. We note that many common AB metrics can be categorized as taking into account particular attributes of clicking behavior. The most common attributes include (i) estimating clicks as indicative of satisfaction or not, (ii) giving particular importance to clicks at the top position of Web search results, and (iii) measuring the time spent by the user prior to clicking. Consequently, we implement the following AB metrics. An overview is given in Table 1.

3.1.1 Click-through Rate

Click-through rate (CTR) is often used as a baseline AB metric, e.g., in [1]. It can be implemented as the average number of clicks per search result page, or as the fraction of pages for which there are any clicks. We follow the second definition. We use $|C^q|$ to denote the number of clicks for query q. The indicator function $\mathbf{1}(\cdot)$ is used and evaluates to 1 when the argument is true, and 0 otherwise.

3.1.2 Click Rank

It was noted by [1] that of all the AB metrics studied in a large scale comparison of AB tests and interleaving evaluation, the AB metric that most reliably agreed with known experimental outcomes was the fraction of search results pages with a click at the top position. As such, we also use two types of metrics: Those which only consider clicks at the top position (named *@1*) and those that consider all clicks (the others). In equations we use $\text{rank}_A(c)$ to denote the rank of click c in the results returned by ranker A.

3.1.3 User Satisfaction

While clicks are often directly interpreted as a user preference, they are known to be both noisy and biased. To remedy the noise, a common approach is to only consider satisfied clicks (here: SAT clicks) with dwell time above a fixed cutoff of 30 seconds [36].

However, only using time as a threshold for satisfaction is problematic as some queries naturally require users to spend more time than others. Recently, Kim et al. [16] showed that taking more user signals into account leads to better prediction of user satisfaction. For this paper, we partially re-implement the SAT click classifier from that work.Our classifier uses the dwell time, document readability, document topic and query topic features suggested by Kim et al. [16]. In particular, the features beyond dwell time are assumed to partially explain the dwell time necessary for a given query and document. We combine these features using quantile regression forests [21]. The model is trained to predict the probability of a SAT click, given user signals. It can be turned into a classifier by selecting a decision threshold, e.g., based on the distribution over classes in the training set. With training on approximately 3,000 manually labeled clicks, our classifier obtains an accuracy of 77%, which is marginally lower than the 81% reported by Kim et al. [16]. The major difference between the implementations is that we do not represent dwell time distributions per topic. Instead, we use raw dwell time values directly as input for our classifier.

The output of this SAT click classifier is used throughout this paper. For a given click c, we define $\text{sat}(c)$ as the estimated probability that c indicates user satisfaction. For succinctness, we also define $\text{is\_sat}(c) := true$ whenever $sat(c) > 0.8$ (the threshold based on the class distribution). Half of the AB metrics we consider use the $\text{is\_sat}(c)$ signal to filter out clicks c that are not deemed satisfied by our classifier. These AB metrics are marked with subscript S.

3.1.4 Time To Click

Another commonly used metric is the time that the user spends on the search result page before clicking a document. As time spent is the key cost to search system users, reducing this time is considered good (e.g. [1]). Our metrics that measure time to click are marked with subscript T. In equations we use $\text{time}(c)$ to denote the time from the user issuing the query until the click c.

Combining all possible choices of AB metrics leads to the eight metrics shown in Table 1. The first four (AB, AB@1, AB_S, AB_S@1) focus on the presence of a click, while the other four capture the time to the first click of a particular type, if such a click occurred (AB_T, AB_T@1, $AB_{T,S}$ and $AB_{T,S}$@1).

3.2 Interleaving

In this paper, we use Team Draft Interleaving (TDI) [24] as our interleaving baseline. This algorithm is most frequently used in

Algorithm 1 Team-Draft Interleaving [24]

1: **Input**: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
2: **Init**: $L \leftarrow (); \sim TeamA \leftarrow \emptyset; \sim TeamB \leftarrow \emptyset; i \leftarrow 1$
3: **while** $A[i] = B[i]$ **do** ..... *common prefix*
4: $L \leftarrow L + A[i]$ ..... *append result to L without assigning teams*
5: $i \leftarrow i + 1$ ..... *increment i*
6: **while** $(\exists i : A[i] \notin L) \wedge (\exists j : B[j] \notin L)$ **do** *not at end of A or B*
7: **if** $(|TeamA| < |TeamB|) \vee$
$((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
8: $k \leftarrow \min_i\{i : A[i] \notin L\}$ ..... *top result in A not yet in L*
9: $L \leftarrow L + A[k]$ ..... *append it to L*
10: $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... *clicks credited to A*
11: **else**
12: $k \leftarrow \min_i\{i : B[i] \notin L\}$ ..... *top result in B not yet in L*
13: $L \leftarrow L + B[k]$ ..... *append it to L*
14: $TeamB \leftarrow TeamB \cup \{B[k]\}$ ..... *clicks credited to B*
15: **Output**: Interleaved ranking $L, TeamA, TeamB$

practice, and has been empirically shown to be equally effective as Balanced Interleaving [1, 14].

Given an incoming user query, TDI produces a result list as follows. The algorithm takes as input two ranked lists of documents for the query, $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$. The goal is to produce a combined ranking $L = (l_1, l_2, \dots)$. This is done in the same way that sports teams may be constructed in a friendly match, with two team captains taking turns picking players for their team.

The algorithm is detailed in Algorithm 1. It initializes the interleaved list L with any common prefix of A and B, if this exists. For this common prefix, no teams are assigned, as no preferences should be inferred.[1] Then, on line 6, the algorithm continues in phases by adding two documents to L: In each phase, on line 7, we first flip an unbiased coin to decide if ranker A or B is given priority. Assuming that ranker A is given priority, A appends its highest ranked result that is not already in L to L (i.e. $l_1 \leftarrow a_1$ in the first instance), and assigns it to $TeamA$. Then, B selects its first result not already present in L (in the first instance either b_1 if it differs from a_1, and b_2 otherwise) and again appends it to L and $TeamB$. This repeats until all results in A or B have been consumed or until L reaches the desired length.

The interleaved ranking L is then shown to the user. Any clicks on documents contributed by A (in $TeamA$) are credited to A. Clicks on documents in $TeamB$ are credited to B. Over an observed sample of interleaving observations, a preference for A or B is then inferred based on which ranker was credited with more clicks. The final outcome of the interleaving comparison experiment can thus be written as:

$$O_{TDI}(A, B) = \text{sign}\left(\frac{1}{|Q|}\sum_{q \in Q} |C_A^q| - |C_B^q|\right), \qquad (1)$$

where Q is the set of all query impressions (non-unique queries issued by all users during the interleaving experiment), and C_X^q denotes the set of clicks observed in $TeamX$ on q.

We use TDI as our baseline throughout this paper. In addition, all our approaches use the *interleaving* (i.e., list construction) algorithm shown in Algorithm 1 and described above. Our focus is on replacing the credit function (Eq., 1). We will introduce our methods in Section 5.

4. DATA ANALYSIS

Many of the AB user satisfaction methods that we introduced in Section 3.1 have been developed recently. Therefore, it is not clear to what degree interleaved comparisons agree with these metrics. In this section, we conduct an empirical analysis of the sensitivity and directional agreement between these AB metrics and TDI. We start by describing the data we use in this section and in the remainder of this paper (4.1). We then propose a new approach for comparing the sensitivity (in terms of statistical power) of AB and TDI comparisons, and use this method to analyze the relative power of the different approaches (4.2). This also lets us estimate the probability of agreement between approaches at varying sample sizes. The results of our analysis are presented in Section 4.3. They motivate why an improved approach is needed, as discussed in Section 4.4.

4.1 Data

For our experiments, we start with a set of 38 pairs of rankers for which both an AB comparison, and a TDI interleaving comparison were performed. All ranker comparisons reflect changes that are typical for the normal development of a commercial web search engine. They consist of changes to the ranking function used to order web search results, in terms of parameters of the ranking function, modified ranking features, and so forth. The comparisons were all run in the first 9 months of 2014, in the United States locale. The experimental unit consisted of assigning users to individual ranking conditions uniformly at random.

The AB and interleaving comparisons were run for varying durations, usually around one week for AB comparisons and around 4 days for interleaving comparisons. Additionally, AB comparisons were typically run with higher volume, resulting in about 80 times more queries for each AB comparison than each interleaving comparison. In all, this data consists of over 3 billion clicks. Depending on the experiment, between 2% and 30% of interleaved queries with clicks had at least one click on a result assigned to one of the teams.

4.2 Estimating Power and Agreement

We now propose an approach for assessing the relative power of AB measurement compared to TDI, and further show how this approach can be used to estimate agreement between approaches at varying sample sizes.

As described earlier, AB tests perform controlled experiments. Users are exposed to either treatment or control result rankings, rendering this a *between subject* experiment. In interleaving, each user is exposed to results from both rankers, rendering them *within subject* experiments. We can measure the importance of this difference using a power computation, which tells us how many independent samples we need to obtain a statistically significant outcome for each approach, as follows.

We start with AB comparisons, following the standard methodology described in [17]: Two independent samples are collected by exposing a random fraction of users to treatment A, and another to treatment B. An AB metric is used to assess each sample, and we are interested in determining whether there is a statistically significant difference between A and B in terms of this metric. This question is typically addressed using a two-sample t-test.

Power of AB comparisons. Assume that the target metric is approximately normally distributed (this is reasonable due to the central limit theorem), with means μ_A and μ_B and equal variance σ_{AB}. Formally, we have $A \sim \mathcal{N}(\mu_A, \sigma_{AB})$ and $B \sim \mathcal{N}(\mu_B, \sigma_{AB})$. We are interested in detecting whether $\mu_A \stackrel{?}{=} \mu_B$. This gives us the null hypothesis $H_0 : \mu_A = \mu_B$ and the alternative hypothesis $H_1 : \mu_A \neq \mu_B$. We also choose the probability of Type-I error we are willing to accept, e.g., $\alpha = 0.05$. The t-test then assumes that H_0 is true and assesses the probability of observing differences of at least the observed sample difference $|\hat{A} - \hat{B}|$ under H_0.

While Type-I error is controlled in the significance test, here we

[1] This was shown to substantially increase sensitivity of the simpler original TDI algorithm [1, 22].

are interested in the *power* (also called sensitivity) of the conducted test. Assuming H_1 is actually true, power quantifies the probability of correctly rejecting H_0. It is affected by the true effect size $\delta_{AB} = (\mu_A - \mu_B)/\sqrt{1/n_A + 1/n_B}\sigma$, where n_A, n_B are the respective sample sizes.

We can assess the power of a test as follows. Under H_1 (samples are drawn from normal distributions with means $\mu_A \neq \mu_B$ and shared variance σ_{AB}) we observe sample A, B and compute the test statistic [19]: $t(A,B) = \frac{(\bar{A}-\bar{B})}{\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} / \sqrt{\frac{\sum(A_i-\bar{A})^2+\sum(B_j-\bar{B})^2}{v_{AB}}}$. The test statistic $t(A,B)$ follows a non-central t distribution $\bar{A} - \bar{B} \sim nct(\delta_{AB}, v_{AB})$, with non-centrality parameter δ_{AB} (the effect size, from above) and degrees of freedom $v_{AB} = n_A + n_B - 2$.

H_0 is correctly rejected when $t(A,B) \geq C_0$. The power of the test is the probability $P(reject(H_0)|H_1) = P(t(A,B) \geq C_0)$ and can be computed (solved using linear programming[2]):

$$P(t(A,B) \geq C_0) = \int_{C_0}^{\infty} nct(\delta_{AB}, v_{AB})dy. \quad (2)$$

Power of interleaving comparisons. The analysis for interleaving is closely related, but relies on the typically more powerful (one-sample) paired t-test. Instead of independent samples, we now observe a single sample I of paired comparisons, assumed to be normally distributed with $I \sim \mathcal{N}(\mu_I, \sigma_I)$. We want to detect whether $\mu_I \stackrel{?}{=} 0$ with $H_0 : \mu_I = 0$ and $H_1 : \mu_I \neq 0$. Given μ_I and sample size n_I, the test statistic $t(I)$ follows a non-central t-distribution $t(I) \sim nct(\delta_I, v_I)$ with non-centrality parameter $\delta_I = \sqrt{n\mu_I/\sigma_I}$ and $v_I = n_I - 1$ degrees of freedom. The power calculation is[3]

$$P(t(I) \geq C_0) = \int_{C_0}^{\infty} nct(\delta_I, v_I)dy. \quad (3)$$

Probability of agreement. Given Equations (2) and (3), we can compute the probability of comparison outcomes at varying sample sizes. For example, the probability that an AB comparison with parameters $\mu_A, \mu_B, \sigma_{AB} : \mu_A - \mu_B > 0$ agrees with the true AB outcome at sample sizes n_A, n_B is computed by plugging into Equation (2) and computing $P(\bar{A} - \bar{B} > 0) = P(t(A,B) > 0)$. Correspondingly, the probability that an interleaving comparison with parameters μ_I, σ_I would agree with the same outcome is computed using Equation (3) so that $P(\bar{I} > 0) = P(t(I) > 0)$.

4.3 Data Analysis Results

We apply the analysis methodology described above to the set of 38 ranking algorithms described in Section 4.1. In Figure 1 we show the power obtained using AB comparisons and interleaving comparisons at increasing sample sizes. We see that on average across the set of 38 experiments, 80% power[4] with AB experiments is obtained with between 10^7 and 10^8 observations (queries). On the other hand, the same power is obtained with between 10^5 and 10^6 observations with TDI. This difference of approximately two orders of magnitude is also consistent with previous work [1].

Having presented the relative power of the approaches, we return to the key question: Do the metrics agree on which ranker is better? We use the method developed in the previous section to estimate agreement between AB and interleaving comparisons, and

[2]We use the python implementation `statsmodels.stats.power.tt_ind_solve_power`, http://statsmodels.sourceforge.net/

[3]We use the python implementation `statsmodels.stats.power.tt_solve_power`.

[4]Controlled experiments are typically designed for 80-95% power.

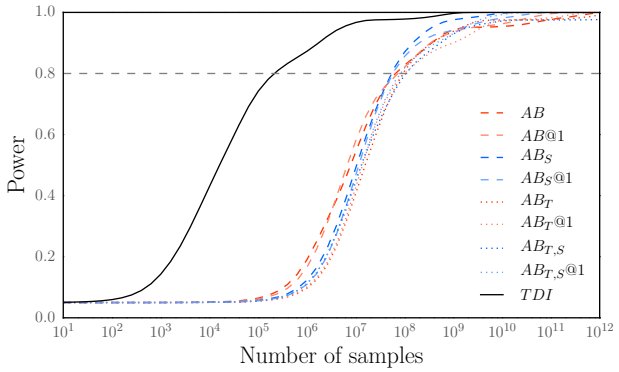

Figure 1: Power as a function of sample size, computed using the observed effect sizes for 38 interleaving and AB comparisons, averaged over all comparisons (assuming two-sided t-test with $p = 0.05$, as described in Section 4.2).

Table 2: Agreement of AB metrics on our data. We measure agreement with TDI, with a sub-sample of AB of the same size as the TDI comparison, and with a sub-sample of AB with the same size as the original AB comparison (this is an upper-bound on agreement for each AB metric). For TDI, values in bold are statically significantly different from 50%.

		self-agreement	
AB Metric	TDI	AB_{Sub}	AB_{Up}
AB	0.63	0.63	0.94
AB@1	**0.71**	0.62	0.95
AB_S	**0.71**	0.61	0.96
AB_S@1	**0.76**	0.60	0.95
AB_T	0.53	0.58	0.91
AB_T@1	0.45	0.59	0.90
$AB_{T,S}$	0.47	0.59	0.88
$AB_{T,S}$@1	0.42	0.60	0.87

AB comparisons at varying sample sizes. The first column in Table 2 summarizes the agreement rates of TDI and AB metrics. We see that agreement rates are generally low. This agreement of TDI with AB metrics is a *baseline* that we want to improve upon in this paper. In the results in Section 6, this baseline is referred to as TDI.

Recall that our goal is to predict the outcome of a large AB comparison given the much smaller amount of data used in interleaving comparisons. As such, another reasonable baseline is to assess how well a smaller AB comparison predicts the outcome of the full AB comparison. We can answer this question using the methodology for assessing the probability of agreement developed above (Section 4.2), by plugging in the observed effect sizes and setting the sample size to that of the corresponding interleaving comparison. This our second *baseline*, which we refer to as AB_{Sub}. Results are given in the second column of Table 2. Generally, these AB metrics computed on small subsamples have low agreement with the experiment outcome given the complete data, of around 60%.

We also compute an upper bound of agreement with AB metrics by actually measuring how well a subsample of the same size as the full AB, AB_{Up}, would agree with itself, using the same methodology. This can be seen as a measure of how predictable a metric is. The last column of Table 2 shows that the time-based metrics are much less predictable than count-based AB metrics. Given these lower and upper bounds on AB agreement with itself, we can restate our goal as follows. We aim to augment TDI to improve over the above two baselines (TDI and AB_{Sub}) and to close the gap with the upper bound AB_{Up}.

Table 3: Definitions for interleaving credit functions. The $\delta(C_A)$ functions give credit to ranker A based on attributes of the clicked documents assigned to ranker A. The last row actually computes a combination of credit functions above it.

Credit functions designed to match AB metrics (cf., 5.2)		$\delta(C_A) =$
TDI	Number of clicks on ranker A	$\lvert C_A \rvert$
TDI@1	Number of clicks on documents that A ranks first	$\sum_{c \in C_A} \mathbf{1}(\text{rank}_A(c) = 1)$
TDI_S	Number of SAT clicked documents contributed by A	$\sum_{c \in C_A} \mathbf{1}(\text{is\_sat}(c))$
TDI_S@1	Number of SAT clicked document ranked first by A	$\sum_{c \in C_A} \mathbf{1}(\text{is\_sat}(c)) \cdot \mathbf{1}(\text{rank}_A(c) = 1)$
TDI_T	Time to clicks on documents contributed by A	$\sum_{c \in C_A} \text{time}(c)$
TDI_T@1	Time to clicks on documents ranked first by A	$\sum_{c \in C_A} \mathbf{1}(\text{rank}_A(c) = 1) \cdot \text{time}(c)$
$TDI_{T,S}$	Time to SAT clicks on documents contributed by A	$\sum_{c \in C_A} \mathbf{1}(\text{is\_sat}(c)) \cdot \text{time}(c)$
$TDI_{T,S}$@1	Time to SAT clicks on documents ranked first by A	$\sum_{c \in C_A} \mathbf{1}(\text{rank}_A(c) = 1) \cdot \mathbf{1}(\text{is\_sat}(c)) \cdot \text{time}(c)$
Parameterized credit functions (cf., 5.3)		$\delta(C_A) =$
$TDI_S^{t_s}, t_s \in \{0.1..0.9\}$	Number of clicks with SAT probability $\geq t_s$, on documents contributed by ranker A	$\sum_{c \in C_A} \mathbf{1}(\text{sat}(c) \geq t_s)$
$TDI_{T,S}^{t_s}, t_s \in \{0.1..0.9\}$	Time to clicks with SAT probability $\geq t_s$, on documents contributed by ranker A $\geq t_s$	$\sum_{c \in C_A} \mathbf{1}(\text{sat}(c) \geq t_s) \cdot \text{time}(c)$
Combined credit functions (cf., 5.4)		$\delta(C_A) =$
$TDI_{T,S}^{\mathbf{w}}, w_i \in \{0.1..0.9\}$	Weighted combination of the credit functions above	$\sum_{w_i \in \mathbf{w}} w_i \delta_i(C_A)$

4.4 Implications

Summarizing the results above motivates the rest of this work: AB metrics have been developed guided by real analysis of user behavior. Yet they usually have relatively low power. Interleaving has much higher power, but low agreement with most AB metrics, being blind to richer behavioral signals. Thus, we aim to optimize interleaving to increase agreement with AB metrics, while maintaining the statistical power of the technique. The AB metric is treated as the gold standard to which interleaving must compare itself.

As noted in the Section 2, this goal is similar to that addressed by [37]. However, they only focused on optimizing sensitivity, while we focus on optimizing correctness in the sense of agreeing with AB metrics. It is similar to [12, 38] in the sense that our method can reduce click bias in interleaved comparisons, however the earlier work only considered agreement with offline metrics.

5. METHOD

In this section we describe how to incorporate user signals into TDI comparisons, to increase agreement with AB metrics. We first formalize the notion of interleaving credit in a way that allows us to incorporate user signals (Section 5.1). We then design a set of credit functions that closely match user satisfaction AB metrics (Section 5.2). Because agreement between interleaving and AB metrics is not necessarily maximized by mirroring AB parameters, we then introduce parameterized credit functions (Section 5.3), and combined credit functions (Section 5.4) designed to be automatically tuned to maximize agreement. Finally, our methodology for maximizing agreement is detailed in Section 5.5.

5.1 Formalizing Interleaving Credit

Formally, for all pairs of rankers A and B, we aim to find an interleaving method that agrees with the *sign* of the differences in AB metrics that we found in an AB comparison. The sign of such a difference should be interpreted as a preference for either A, B, or neither. We denote such a preference, the comparison outcome O_{AB}, of the metric AB as:

$$O_{AB}(A, B) = \text{sign}(AB(A) - AB(B)). \quad (4)$$

Instances of AB metrics are *click through rate*, *clicks at one*, and *time to click* (see Table 1 for details).

As opposed to AB metrics, interleaving methods are directly defined on pairs of rankers. Following the same notation, the outcome of an interleaving comparison with TDI can thus be denoted as:

$$O_{TDI}(A, B) = \text{sign}(TDI(A, B)). \quad (5)$$

Interleaving preferences, when using TDI (cf., Section 3.2), come from differences between credit acquired by each ranker:

$$TDI(A, B) = \frac{1}{|Q|} \sum_{q \in Q} \delta(C_A^q) - \delta(C_B^q). \quad (6)$$

Here, Q is a set of query impressions and $\delta(C_A)$ a credit function that attributes credit to ranker A depending on user interactions with the result list. Next, we introduce a new set of credit functions that is designed to mirror the use of user signals in AB comparisons.

5.2 Matching AB Credit

We now present instantiations of credit functions $\delta(C_A)$ designed to match the AB metrics in Section 3.1. All our interleaving credit functions are defined on a set of clicks assigned to a ranker (e.g., for ranker A these are $c \in C_A$), for a query impression. Clicks are associated with user signals.

The details of our *matching* credit functions are given in the first part of Table 3. The following signals are used:

- $\lvert C_A \rvert$ the number of clicks for ranker A for a query impression. See Section 3.1.1.
- $\text{rank}_A(c)$ is the rank of the clicked document in the original ranking A (before interleaving: i.e., rankers A and B can have different documents at rank 1). See Section 3.1.2 for the a description of this signal as used in AB metrics.
- $\text{is\_sat}(c)$ is a binary indicator that is *true* if the SAT classifier identified the click as SAT click, see Section 3.1.3 for details on the SAT click classifier.
- $\text{sat}(c)$ is the probability of the click being a SAT click. Again, details are in Section 3.1.3.

- $\text{time}(c)$ is the time from query submission to the observed click, in seconds. See Section 3.1.4 for the corresponding AB signal.

Previously proposed interleaved comparison methods, such as TDI, use the credit function TDI shown in the table. It can be interpreted as a close match to the AB metric AB, because it estimates whether a given ranker would have obtained a click in an AB comparison.

The credit functions $TDI@1$, TDI_S, and $TDI_S@1$ are designed to closely match the AB metrics $AB@1$, AB_S, and $AB_S@1$. For clicks at rank one, we consider whether a clicked document would have been placed first by the original ranker, as this reflects the most accurately whether the ranker would be likely to receive a click at the top rank in an AB comparison. For SAT clicks, we use the same classifier as for our AB comparisons above, as in Section 3.1.3.

The four time-based credit functions TDI_T, $TDI_T@1$, $TDI_{T,S}$, and $TDI_{T,S}@1$ are designed to match the time-based AB metrics. E.g., TDI_T reflects the AB metric AB_T, however we use the average time to click for a ranker, as it tends to be more robust than the time to the first click. The remaining three metrics implement filters on which clicks contribute, parallel to the click-based metrics described above.

5.3 Parameterized Credit Functions

Next, we propose a second set of interleaving credit functions that can be parameterized to automatically calibrate them to maximize agreement with AB metrics. Effectively calibrating these credit functions would allow users of interleaved comparisons to automatically identify credit functions that maximize agreement with arbitrary AB metrics.

For instance, we define a credit function that captures user satisfaction. We filter out clicks c that have a low satisfaction probability $\text{sat}(c)$ by thresholding this probability using a threshold t_s. This leads to the following credit function:

$$\delta(C_A)^{t_s} = \sum_{c \in C_A} \mathbf{1}(\text{sat}(c) > t_s). \quad (7)$$

The threshold, t_s in this case, of such a parametrized credit function can be tuned to maximize agreement with AB metrics. We define two such parameterized functions, the first click-based, as shown above, the second time-based. We list our parameterized credit functions in the second part of Table 3.

5.4 Combined Credit Functions

Now that we have several credit functions, as listed in the first two sections of Table 3, we can take it a step further and start combining them. We propose to combine the interleaving credit functions in a weighted linear combination:

$$TDI^{\mathbf{w}}(A,B) = \frac{1}{|Q|} \sum_{q \in Q} \sum_{w_i \in \mathbf{w}} w_i \delta_i(C_A^q) - w_i \delta_i(C_B^q), \quad (8)$$

where $\mathbf{w}$ denotes the weights used to combine several credit functions. We thus define the interleaving preference as a weighted sum of credit functions we introduced earlier.

In the original TDI, we have a single credit function as defined in the first row of Table 3 and a weight vector of $\mathbf{w} = (1)$.

5.5 Maximizing Agreement with AB Metrics

We return to our initial goal, to optimize the agreement between interleaving metrics and AB metrics, and present a method for automatically tuning interleaving credit functions to maximize agreement with a given AB metric. Together with the parameterized and combined credit functions presented above, this allows tuning interleaving to an arbitrary AB metric. Our approach treats the AB metric as a black box presented by an experimenter who presumably selected this metric for some reason.

Algorithm 2 Maximizing Agreement

1: **Input**: Ranker pairs $C = ((A_1, B_1), \ldots, (A_n, B_n))$, AB
2: **Init**: test agreements $A \leftarrow []$, weights $W \leftarrow []$
3: **for all** $n \leq N$ **do** ..... *N repetitions*
4: $S \leftarrow sample\_with\_rep(C, |C|)$. *bootstrap sample, train set*
5: $\hat{\mathbf{w}} \leftarrow \arg\max_{\mathbf{w}} \sum_{(A,B) \in S} \mathbf{1}(O^{\mathbf{w}}_{TDI}(A,B) = O_{AB}(A,B))$
6: $O \leftarrow C \setminus S$ ..... *'out of bag' sample, test set*
7: $A \leftarrow A + \frac{1}{|O|} \sum_{(A,B) \in O} \mathbf{1}(O^{\hat{\mathbf{w}}}_{TDI}(A,B) = O_{AB}(A,B))$
8: $W \leftarrow W + \hat{\mathbf{w}}$ ..... *append weight vector*
9: **Output**: weights $mean(W)$, agreement $mean(A)$

The weights introduced in Equation (8) can be optimized such that we maximize the agreement:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{(A,B) \in S} \mathbf{1}(O^{\mathbf{w}}_{TDI}(A,B) = O_{AB}(A,B)). \quad (9)$$

I.e., we maximize the number of times the outcome of an AB comparison agrees with the outcome of an interleaving comparison for all of ranker comparisons S.

To implement and validate the maximization step in Equation (9), we use the bootstrap procedure presented in Algorithm 2. This takes as input a set C of pairs of rankers that have been compared (such as those described in described in Section 4.1) and an AB metric such as AB. For N repetitions, a bootstrap sample S of size $|C|$ is taken from C. On this sample we compute $\hat{\mathbf{w}}$ using Equation (9) for all $\mathbf{w}$ we consider. We validate the agreement that this weight vector $\hat{\mathbf{w}}$ gives on unseen data and we report the mean $\hat{\mathbf{w}}$ and mean agreement. In our experiments $N = 100$ and we consider $\mathbf{w} = (w_i \in \{0, 0.01 \ldots 1\}, \ldots, w_n)$.

We use the same procedure to optimize the parameters t_s of the parameterized credit functions described in Section 5.3. Instead of computing the argmax on Line 5 over all $\mathbf{w}$, we compute the optimal $\hat{t_s}$:

$$\hat{t_s} = \arg\max_{t_s} \sum_{(A,B) \in S} \mathbf{1}(O^{t_s}_{TDI}(A,B) = O_{AB}(A,B)). \quad (10)$$

6. EXPERIMENTS AND RESULTS

In Section 4, we examined the agreement between TDI and AB metrics, and the sensitivity of both types of comparison methods. Depending on the AB metric, agreement ranges from random up to 75%, while sensitivity of TDI is on average two orders of magnitude higher than that of AB metrics. In this section we evaluate our new interleaving credit functions. First, we analyze what level of agreement can be reached by matching interleaving credit functions with the parameters of AB metrics. Second, we evaluate our parameterized credit functions, and our method for optimizing agreement with AB metrics.

6.1 Matching AB Credit

In our first set of experiments, we evaluate our *matching* credit functions. These are designed to match the parameters of the AB metric that we wish to optimize, as explained in Section 5.2. For instance, for the target AB metric AB_S we classify observed clicks on interleaving impressions using the same classifier used by the AB comparison, and only assign interleaving credit for satisfied clicks.

Table 4: Agreement of *matching* interleaving credit functions (designed to match AB metric parameters). Boldface indicates values significantly different from 0.5 (two-sided binomial test, p = 0.05). On the diagonal are metrics for which parameters are designed to match (gray background). Best agreement per AB metric is underlined.

	Interleaving Credit							
AB Metric	TDI	TDI@1	TDI_S	TDI_S@1	TDI_T	TDI_T@1	$TDI_{T,S}$	$TDI_{T,S}$@1
AB	0.63	0.66	**0.84**	0.66	0.61	0.61	0.58	0.53
AB@1	**0.71**	**0.68**	**0.76**	0.63	0.63	0.47	0.55	0.55
AB_S	**0.71**	**0.68**	**0.87**	**0.68**	**0.68**	0.58	0.61	0.55
AB_S@1	**0.76**	**0.68**	**0.82**	0.63	**0.74**	0.53	0.61	0.50
AB_T	0.53	0.55	0.47	0.55	**0.71**	0.55	**0.68**	0.58
AB_T@1	0.45	0.47	0.45	0.58	0.63	0.58	0.61	0.61
$AB_{T,S}$	0.47	0.55	0.53	**0.71**	0.66	0.66	0.58	0.53
$AB_{T,S}$@1	0.42	0.50	0.53	0.66	0.61	0.66	0.58	0.58

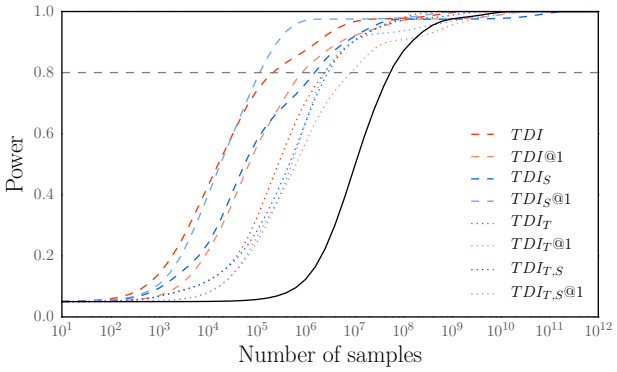

Figure 2: Power for TDI with matching credit functions (assuming two-sided t-test with $p = 0.05$, as described in Section 4.2). The black line denotes AB_S, the AB metric with most power.

As we study 8 AB metrics, this gives rise to 8 possible variants of TDI with matching credit functions.

Table 4 shows the agreement between each AB metric and each variant of TDI. In the first column, we see the agreement between baseline TDI and each AB metric, computing each as previously defined. The lowest agreement is observed between the original TDI and the AB metric $AB_{T,S}$@1, at 42%. The highest agreement is observed between TDI_S and the AB metric it is designed to optimize (AB_S), with 87%. Given the small sample of 38 comparisons, only the agreement rates above 68% are statistically significantly different from random agreement, and are shown in bold in the table. These compare favorably with typical inter-judge agreement rates in offline evaluations of around 65% [32], and with the bounds on AB self-agreement AB_{Sub}, AB_{Up} in Table 2.

We note that using different credit functions often increases agreement between AB metrics and TDI, but interestingly the maximal agreement is often not seen when the AB metric matches the credit function used for interleaving. This can be observed by comparing the metrics that match in terms of their parameters (indicated by the gray cells in Table 4), to the ones that achieved highest agreement (underlined). For example, agreement with AB_S is maximized by TDI_S, but agreement with $AB_{T,S}$ is maximized by TDI_S@1. A reason for this is the interplay between bias and noise. By more aggressively removing noise in the interleaving comparison (in this case, by only considering SAT clicks at the top position), we may increase agreement with related AB metrics, even those for which there is bias due to a slight mismatch between the interleaving and AB metric.

Our results show that agreement between interleaving and AB comparisons can be substantially improved by matching interleaving credit parameters to those of the target AB metrics. We also need to ensure that in doing so, we do not decrease the sensitivity of interleaving. Intuitively, removing observations (e.g., clicks beyond the first position) may reduce sensitivity. On the other hand, if the removed observations are noisy, the interleaving signal may actually become more discriminative, and sensitivity can be increased. Figure 2 shows the power for TDI with replaced scoring functions. We see that TDI with matching credit functions typically has lower power than standard TDI. In particular, sensitivity decreases for time-based metrics, which may also explain the relatively lower agreement between time-based interleaving credit functions and AB metrics. However, the power of these variants of TDI is still 1 to 2 orders of magnitude larger than the power of the AB metric with most power. Sensitivity is increased by TDI_S, the credit function that also shows highest agreement. This result indicates that focusing interleaving credit on low-noise clicks is a very promising way to achieve both high sensitivity and good agreement with user satisfaction metrics.

The results of the analysis in this section motivate the next set of questions. Given a target AB metric, what is the best credit function that should be applied to TDI? Just as the correct credit function may not be the same as the target AB metric, the parameters of the credit function may need to be tuned. And, once we automatically optimize the parameters of interleaving credit functions, to what degree do optimal values generalize to unseen ranker comparisons? We address these questions next.

6.2 Parameterized Credit Functions

One way to increase agreement of TDI with AB metrics is to take an interleaving credit function with parameters (see Section 5.3) and tune the parameters towards a given AB metric. For instance, previous work has shown that it is possible to estimate the probability that a given click indicates user satisfaction [16]. While an AB metric such as AB_S must incorporate a threshold below which clicks are not considered to indicate user satisfaction, the threshold for TDI need not be the same. Rather, we can find the optimal threshold t_s for $TDI_S^{t_s}$ at which to consider a click as satisfied. This optimization procedure might lead to reduced variance, and thereby increase agreement with AB metrics.

We use the maximization procedure described in Section 5.5 and in particular Equation (10) to find an optimal threshold for each AB metric we consider. Note that, as opposed to experiments in the previous section, here we obtain averages over $N = 100$ iterations of the maximization procedure, instead of averages over the 38 comparisons. This allows us to perform statistical significance testing using a one-sample two-sided student's t-test. In our result table we indicate statistically significant improvements over TDI by $^{\blacktriangle}$ ($p < 0.01$) (losses $^{\blacktriangledown}$). Also, as opposed to before, we now

Table 5: Agreement for $TDI_S^{t_s}$, $TDI_{T,S}^{t_s}$, and $TDI_{T,S}^{\mathbf{w}}$. Parameters t_s and $\mathbf{w}$ are chosen to maximize agreement with the AB metrics on held out data. Higher agreement than TDI is bold. Stat. sig. improvements over TDI are indicated by $^\blacktriangle$ ($p < 0.01$) (losses $^\blacktriangledown$).

		(a) $TDI_S^{t_s}$		(b) $TDI_{T,S}^{t_s}$		(c) $TDI_{T,S}^{\mathbf{w}}$		
AB Metric	TDI	Agreement	t_s	Agreement	t_s	Agreement	w_1	w_2
AB	0.63	**0.82** $^\blacktriangle$	0.76 (0.09)	0.53 $^\blacktriangledown$	0.52 (0.40)	**0.84** $^\blacktriangle$	1.00 (0.00)	0.00 (0.00)
AB@1	0.71	**0.79** $^\blacktriangle$	0.74 (0.19)	0.54 $^\blacktriangledown$	0.40 (0.32)	**0.75** $^\blacktriangle$	1.00 (0.00)	0.05 (0.22)
AB_S	0.71	**0.84** $^\blacktriangle$	0.76 (0.09)	0.48 $^\blacktriangledown$	0.29 (0.31)	**0.85** $^\blacktriangle$	1.00 (0.00)	0.00 (0.00)
AB_S@1	0.76	**0.84** $^\blacktriangle$	0.68 (0.24)	0.48 $^\blacktriangledown$	0.37 (0.32)	**0.82** $^\blacktriangle$	1.00 (0.00)	0.02 (0.14)
AB_T	0.53	0.47 $^\blacktriangledown$	0.67 (0.28)	**0.67** $^\blacktriangle$	0.54 (0.27)	**0.68** $^\blacktriangle$	0.99 (0.11)	0.90 (0.30)
AB_T@1	0.45	**0.49** $^\blacktriangle$	0.57 (0.35)	**0.62** $^\blacktriangle$	0.61 (0.34)	**0.56** $^\blacktriangle$	0.96 (0.22)	0.79 (0.41)
$AB_{T,S}$	0.47	0.46	0.46 (0.38)	**0.61** $^\blacktriangle$	0.41 (0.30)	**0.63** $^\blacktriangle$	0.91 (0.30)	0.88 (0.33)
$AB_{T,S}$@1	0.42	**0.52** $^\blacktriangle$	0.30 (0.39)	**0.62** $^\blacktriangle$	0.42 (0.34)	**0.50** $^\blacktriangle$	0.06 (0.65)	0.25 (0.41)

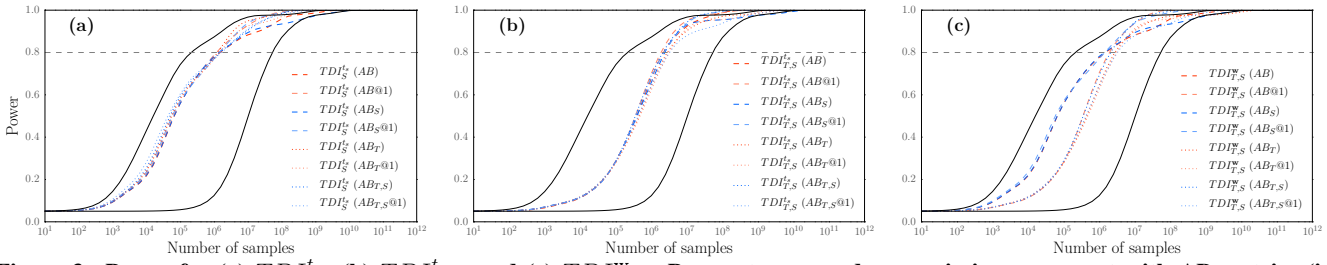

Figure 3: Power for (a) $TDI_S^{t_s}$, (b) $TDI_{T,S}^{t_s}$, and (c) $TDI_{T,S}^{\mathbf{w}}$. Parameters t_s and $\mathbf{w}$ maximize agreement with AB metrics (in brackets) on held out data, see Table 5. The upper black line denotes TDI, the lower AB_S, the AB metric with most power.

measure how far our optimized credit functions generalize to unseen data.

The results for maximizing agreement of $TDI_S^{t_s}$ are shown in column (a) in Table 5. In the table, we see substantially and signficantly increased agreement rates of up to 84% for the AB metrics that only depend on clicks, reducing disagreement rates by between 6% and 20%. E.g. in Table 5, for $t_s = 0.76$ in the first row means that clicks predicted to have less than a 76% chance of indicating satisfaction are ignored. This causes 58% of clicks to be ignored on average. We see that across many folds, the optimal value is between 0.67 and 0.85, and we found that the precise value does not impact the outcome significantly.

Interestingly, the optimal threshold at which clicks should be included in the score calculation are around $t_s = 0.75$, which is lower than the $t_s = 0.8$ which is used in the AB_S and AB_S@1 AB metrics. It is equally interesting that learning such a threshold changes agreement between TDI and time-based AB metrics much less but generally decreases it, and selects a t_s threshold that is much lower.

In contrast, if we take the credit function $TDI_{T,S}^{t_s}$, that does take time as well as satisfaction into account and if we learn the same threshold t_s, then in column (b) of Table 5 we see a very different result than before in column (a). Now, the agreement between TDI and AB metrics that incorporate time increases substantially and significantly (from 42-53% to 61-67%, a net disagreement error reduction of between 4% and 22%) while disagreement increases significantly with the non-time based metrics. Now, for the time based metrics we also outperform the baseline AB_{Sub} which uses as many query impressions as TDI does (see Table 2).

These changes in agreement exhibit the same pattern as seen when tuning a simpler threshold on satisfaction (as reported in column (a) in Table 5). In particular, the deterioration in agreement from tuning a feature that does not represent a measure included in the AB metric reduces agreement. We hypothesize that this is due to the maximization procedure failing to find a optimal value that generalizes well to unseen data, as the target AB metrics that are based only on clicks have low correlation with the credit function we are optimizing. These results motivate our next approach: optimize a combination of interleaving credit functions that best matches a given AB metric.

But first, we look at what happens to the power of TDI when we optimize parameters of a credit function. Results are in Figures 3 (a) and 3 (b). We see that again, the power for our adjust interleaving credit functions lies between standard TDI and the AB metric with the highest power. In other words, we increased agreement while maintaining an advantage in terms of power over AB comparisons.

6.3 Combined Credit Functions

As we saw in the previous section and in columns (a) and (b) in Table 5, for different types of AB metrics we need different interleaving credit functions to increase agreement. Optimizing a single parameter (t_s) for a single credit function proved not powerful enough. In this section we use the maximization procedure described in Section 5.5 and in particular Equation (9) to find weights $\mathbf{w}$ for a weighted combination of already optimized credit functions that maximizes agreement. That is, we take for each AB metric the threshold t_s that in the previous section maximized agreement. Note that we only optimize a weighted combination of *two* credit functions, namely, we learn w_1 for $TDI_S^{\hat{t_s}}$ and w_2 $TDI_{T,S}^{\hat{t_s}}$. The intuition behind this simple model is that it should be able to capture attributes of each of the AB metrics.

We obtain the results presented in column (c) in Table 5. Where in the previous section, in columns (a) and (b) in Table 5, we obtained average agreement of 65% and 57% respectively, now we obtain an average agreement of 70%. Agreement with all individual AB metrics increased significantly from 42-76% to 50-85%. Interestingly, we see that the weights w_1 and w_2 that are the result of the optimization procedure are mostly selecting ($w_2 \approx 0$) the $TDI_S^{\hat{t_s}}$ credit function for the click based AB metrics. While the time based AB metrics additionally put weight ($w_2 \gg 0$) on the time based credit function $TDI_{T,S}^{\hat{t_s}}$.

Lastly, turning to the sensitivity, in Figure 3 (c) we see that also for the combined credit functions sensitivity stayed 1 to 2 orders of magnitude higher compared to AB metrics.

7. CONCLUSION

In this paper, we showed how to optimize interleaving outcomes to agree better with a given target AB metric, while maintaining the sensitivity advantage of interleaved comparisons over AB tests. We started by analyzing the agreement of team draft interleaving (TDI) with a set of 8 AB metrics based on combinations of click count, click positions, satisfied clicks, and time to click signals. To enable this analysis, we introduced a method for comparing AB and interleaved comparison metrics in terms of power and agreement across varying sample sizes. We found that, while TDI is very sensitive, its agreement with user satisfaction AB metrics on realistic ranking evaluations is low, from random up to 76%.

Results of this analysis motivated our approach. We proposed to replace the default credit function of TDI with novel credit functions that take richer user signals into account. In particular, we designed sets of credit functions that (1) match the parameters of AB metrics, (2) are parameterized and (3) combine (parameterized) credit functions. To automatically tune the parameters of these last credit functions, we further introduced a bootstrapping algorithm that can automatically maximize agreement with arbitrary AB metrics.

Our empirical results, obtained on 38 paired experiments with a total of 3 billion clicks, showed that our approach can substantially and significantly increase agreement with AB metrics. In particular, learning a combination of parameterized credit functions resulted in agreement of up to 85%, improving the agreement with AB metrics by up to 22% (almost halving these disagreements). We also showed that the sensitivity for all our adapted versions of TDI is still 1 to 2 orders of magnitude higher than that of AB metrics.

The most important implication of our results is that it enables, for the first time, the integration of rich user satisfaction signals with highly sensitive interleaved comparison methods. This will dramatically reduce the required sample sizes, and therefore cost, of such online evaluations. Opportunities for future work include the development of yet more sophisticated (learned) credit functions, e.g., to take into account session-level or task-level features. Furthermore, more accurate prediction of AB outcomes may be possible by additionally taking into account the magnitude and uncertainty of AB metrics of individual experiments. Finally, it would be interesting to see where disagreements between interleaving and AB outcomes are highest; for what queries and for what kind of comparisons.

Acknowledgements. We would like to thank Ahmed Hassan Awadallah, Dan Liebling, Maarten de Rijke, Milad Shokouhi, Nick Craswell, Paul Bennett, Peter Bailey, Rich Caruana, Ryen White for helpful discussions and insights. This research was partially supported by the Dutch national program COMMIT.

REFERENCES

[1] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM TOIS*, 30(1):1–41, Feb. 2012.

[2] F. Diaz, R. White, G. Buscher, and D. Liebling. Robust models of mouse movement on dynamic web search results pages. In *CIKM*, pages 1451–1460. ACM Press, Oct. 2013.

[3] G. Dupret and M. Lalmas. Absence time and user engagement: Evaluating ranking functions. In *WSDM*, pages 173–182, 2013.

[4] S. Fox, K. Karnawat, M. Mydland, S.Dumais, and T.White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23: 147–168, 2005.

[5] Q. Guo and E. Agichtein. Understanding "Abandoned" Ads: Towards Personalized Commercial Intent Inference via Mouse Movement Analysis. *SIGIR-IRA*, 2008.

[6] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI EA*, page 3601, Apr. 2010.

[7] A. Hassan, R. Jones, and K. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM*, 2010.

[8] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM*, 2013.

[9] Y. He and K. Wang. Inferring search behaviors using partially observable markov model with duration (POMD). In *WSDM*, 2011.

[10] W. Hersh, A. H. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR*, pages 17–24, 2000.

[11] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *CIKM*, page 249, 2011.

[12] K. Hofmann, F. Behr, and F. Radlinski. On Caption Bias in Interleaving Experiments. In *CIKM*, 2012.

[13] J. Huang, T. Lin, and R. White. No search result left behind. In *WSDM*, page 203, 2012.

[14] T. Joachims. Evaluating Retrieval Performance using Clickthrough Data. In J. Franke, G. Nakhaeizadeh, and I. Renz, editors, *Text Mining*, pages 79–96. Physica/Springer, 2003.

[15] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM TOIS*, 25(2):7+, Apr. 2007.

[16] Y. Kim, A. Hassan, R. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM*, 2014.

[17] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[18] D. Lagun, C. Hsieh, D. Webster, and V. Navalpakkam. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *SIGIR*, 2014.

[19] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. springer, 2006.

[20] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR '09*, pages 43–50, 2009.

[21] N. Meinshausen. Quantile regression forests. *jmlr*, 7:983–999, 2006.

[22] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR*, pages 667–674, 2010.

[23] F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In *WSDM*, 2013.

[24] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM*, pages 43–52. ACM Press, 2008.

[25] M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[26] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved Comparisons for Fast Online Evaluation. In *CIKM*, pages 71–80, 2014.

[27] A. Schuth, R.-J. Bruintjes, F. Büttner, J. van Doorn, C. Groenland, H. Oosterhuis, C.-N. Tran, B. Veeling, J. van der Velde, R. Wechsler, D. Woudenberg, and M. de Rijke. Probabilistic multileave for online retrieval evaluation. In *SIGIR*, 2015.

[28] Y. Song, X. Shi, R. White, and A. Hassan. Context-Aware Web Search Abandonment Prediction. In *SIGIR*, 2014.

[29] J. Teevan, S. Dumais, and E. Horvitz. The potential value of personalizing search. In *SIGIR*, pages 756–757, 2007.

[30] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR*, pages 225–231, 2001.

[31] A. Turpin and F. Scholar. User performance versus precision measures for simple search tasks. In *SIGIR*, pages 11–18, 2006.

[32] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM*, 36(5):697 – 716, 2000.

[33] H. Wang, Y. Song, M. Chang, X. He, A. Hassan, and R. White. Modeling Action-level Satisfaction for Search Task Satisfaction Prediction. In *SIGIR*, 2014.

[34] K. Wang, T. Walker, and Z. Zheng. PSkip: Estimating relevance ranking quality from web search clickthrough data. In *KDD*, pages 1355–1364, 2009.

[35] K. Wang, N. Gloy, and X. Li. Inferring search behaviors using partially observable Markov (POM) model. In *WSDM*, 2010.

[36] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and Effort: An Analysis of Document Utility. In *CIKM*, 2014.

[37] Y. Yue, Y. Gao, O. Chapelle, Y. Zhang, and T. Joachims. Learning more powerful test statistics for click-based retrieval evaluation. In *SIGIR*, pages 507–514, 2010.

[38] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *WWW*, pages 1011–1018, 2010.