A Reference-Dependent Model for Web Search Evaluation

Understanding and Measuring the Experience of Boundedly Rational Users

Nuo Chen
pleviumtan@toki.waseda.jp
Waseda University
Tokyo, Japan

Jiqun Liu
jiqunliu@ou.edu
University of Oklahoma
Norman, OK, USA

Tetsuya Sakai
tetsuyasakai@acm.org
Waseda University
Tokyo, Japan

ABSTRACT

Previous researches demonstrate that users' actions in search interaction are associated with relative gains and losses to reference points, known as the *reference dependence* effect. However, this widely confirmed effect is not represented in most user models underpinning existing search evaluation metrics. In this study, we propose a new evaluation metric framework, namely *Reference Dependent Metric* (ReDeM), for assessing query-level search by incorporating the effect of reference dependence into the modelling of user search behavior. To test the overall effectiveness of the proposed framework, (1) we evaluate the performance, in terms of correlation with user satisfaction, of ReDeMs built upon different reference points against that of the widely-used metrics on three search datasets; (2) we examine the performance of ReDeMs under different task states, like task difficulty and task urgency; and (3) we analyze the statistical reliability of ReDeMs in terms of *discriminative power*. Experimental results indicate that: (1) ReDeMs integrated with a proper reference point achieve better correlations with user satisfaction than most of the existing metrics, like Discounted Cumulative Gain (DCG) and Rank-Biased Precision (RBP), even though their parameters have already been well-tuned; (2) ReDeMs reach relatively better performance compared to existing metrics when the task triggers a high-level cognitive load; (3) the discriminative power of ReDeMs is far stronger than Expected Reciprocal Rank (ERR), slightly stronger than Precision and similar to DCG, RBP and INST. To our knowledge, this study is the first to explicitly incorporate the reference dependence effect into the user browsing model and offline evaluation metrics. Our work illustrates a promising approach to leveraging the insights about user biases from cognitive psychology in better evaluating user search experience and enhancing user models.

CCS CONCEPTS

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

KEYWORDS

Reference Dependence Effect; cognitive bias; information retrieval; web search evaluation

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583551

ACM Reference Format:
Nuo Chen, Jiqun Liu, and Tetsuya Sakai. 2023. A Reference-Dependent Model for Web Search Evaluation: Understanding and Measuring the Experience of Boundedly Rational Users. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583551

1 INTRODUCTION

Understanding how users think, behave, and make decisions in search interaction attaches substantial importance to developing effective evaluation metrics for information retrieval (IR) systems. To reflect the process through which users interact with search engines, existing offline evaluation metrics often involve a variety of premises about users' behavior. For instance, Rank-Biased Precision (RBP)@ϕ assumes that a user will perpetually examine ranked results, but with a fixed probability of $1 - p$ to leave at each rank [34]. The C/W/L/A framework [32] summarizes the browsing behavior models of most existing metrics by three interrelated functions: (1) Continuation probability (C), (2) Weight function (W) and (3) Last probability (L), as well as a function called *aggregation* (A) that decides how the utility gained by a user from the 1st document to the i-th document is aggregated.

Existing assumptions underpinning most formal models usually define users as rational decision-makers seeking to pursue clear maximized gains. This helps simplify the process of training models and setting parameters. However, the rational user assumption has been challenged by behavioral economic studies arguing that people are *boundedly rational* and their decision-making processes under uncertainty can be influenced by cognitive biases. As a result, people's actual behavior often deviates from what is expected or predicted under rational normative models [50–52]. As a user's search interaction often involves a series of local decision-making processes (e.g., judging the relevance of an item, search stopping, satisfaction feedback) under uncertainty, there is substantial evidence showing that cognitive biases also take place in search interaction [25].

Reference dependence is one of the cognitive biases examined in behavioral economic researches [20, 48]. Evidence from behavioral economics shows that the gain or loss a decision maker perceived is relative to a reference point, rather than being an absolute value [22, 51]. In the field of interactive information retrieval (IIR), Liu and Han [27] investigated the effects of reference points on users' search behavior and satisfaction and showed that users' search behavior and satisfaction are significantly associated with the relative gains and losses to certain reference points. However, the above research did not further incorporate the reference dependence effect into a formal user model of interactive search, leaving a vacuum at the topic of how to develop a more realistic, psychology-informed user model for offline evaluation metrics by taking the reference dependence effect into consideration.

To address the gap discussed above, we propose a new evaluation metric framework that incorporates the effect of reference dependence into user browsing models (refer to Section 3). With a new user model incorporating the reference dependence effect, we develop a series of Rreference-Dependent Metrics (ReDeMs) via the C/W/L/A framework and meta-evaluate their performance in terms of capturing user satisfaction and statistical reliability. Specifically, we seek to answer three research questions based upon experiments on publicly available search datasets (refer to Section 4).

- **RQ1**: Compared to existing metrics, will the ReDeMs have a better correlation with user satisfaction feedbacks?
- **RQ2**: To what extent does the performance of ReDeMs vary under different task states, like task difficulty and task urgency?
- **RQ3**: Compared to existing metrics, how reliable are ReDeMs in terms of *discriminative power* [37]?

RQ1 focuses on the overall performance of the proposed framework, RQ2 focuses on the stability and consistency of metric performance under different task states, RQ3 focuses on the statistical reliability of the proposed metrics.

Experiment results show that: (1) the proposed ReDeMs with a proper reference point can achieve higher correlations with users' satisfaction feedbacks than most of the existing metrics like Precision, Discounted Cumulative Gain (DCG) [19], Rank-Biased Precision (RBP) [35] and INST [5]. Under some occasions, ReDeMs can also achieve higher correlations with users' satisfaction feedbacks than Expected Reciprocal Rank (ERR) [13]. (2) ReDeMs have a generally stable performance under different task states, and have a relatively better performance compared to existing metrics when the task needs a high-level cognitive load. (3) The discriminative power of ReDeMs is far stronger than ERR, slightly stronger than Precision and similar to DCG, RBP and INST.

The main contributions of this paper are as follows:

- Our study explicitly incorporates reference dependence effect into offline evaluation metrics and illustrates a viable approach to developing *user-bias-aware* IR evaluation. Our findings suggest that when constructing an evaluation metric, the relative gain to a reference point should also be considered besides widely used factors, such as absolute final information gain and search efforts.
- Evaluation results demonstrate that the proposed reference dependent metrics (ReDeMs) are effective in capturing users' levels of satisfaction. The ReDeMs were tested in a broad range of topics, task states and search scenarios, and can be implemented and replicated in other IR evaluation contexts with similar judgment labels and search interaction signals.
- The general framework of ReDeMs is consistent with that of existing metrics (e.g., C/W/L/A) and could be easily extended to represent and estimate the impacts of other user biases and situational factors in search evaluation.
- In general, with the growing research attention focusing on algorithmic biases and fairness, our research empirically demonstrates the value of the knowledge regarding *user biases* in evaluating search system performances from user perspective and may inspire more future research on enhancing user-bias-aware search evaluation.

2 PRIOR ART

Evaluating the effectiveness of search engines has long been a central concern for the information retrieval (IR) community. Existing evaluation methods can be broadly divided into two classes, *user-based* (or online) evaluation and *test collection-based* (or offline) evaluation [53].

2.1 Offline Evaluation Metrics and User Models

Offline evaluation is often built upon different simulations of the process of a user interacting with a system under operational settings [44], and the evaluation metric scores can be viewed as the simulation of the *gain* a user accumulated during that process. Based on this basic setup, a range of evaluation metrics involving explicit or implicit user behaviour models were proposed and empirically tested, including Discounted Cumulative Gain (DCG) and its variants [9, 19, 35], Rank-Biased Precision (RBP) [35], Expected Reciprocal Rank (ERR) [13], Expected Browsing Utility (EBU) [57], Time-Biased Gain (TBG) [47], U Measure [41], Inverse Square (INSQ) [33], INST [5], Bejeweled Player Model (BPM) metrics [58], Information Foraging Theory (IFT) measure [4], and so forth. Recently, more IR metrics have been developed for evaluating multi-queries search sessions or estimating user preferences with diversified intentions, but these are out of the scope of this paper. Despite manifold metrics with multifarious models, the user model behind a metric can be deconstructed into three interrelated aspects of the user behavior:

- *Continuation probability*, $C(i)$: the probability that a user who has inspected the i-th item in the SERP will continue to examine the item at rank $i+1$.
- *Weight function*, $W(i)$: the fraction of user attention on the item at position i. In other words, it is the likelihood of a user viewing the item at position i at any time under a sequence of random selections.
- *Last probability*, $L(i)$: the probability that a user examine the document at rank i and then stop interacting with the SERP.

This analytical framework, known as the *C/W/L* framework [31, 34] provide a common ground for comparing models of different metrics. In the C/W/L framework, as long as one of the three components is known, then the other two components can be calculated as well. For example, the $L(i)$ can be calculated by $C(i)$ through:

$$L(i) = (1 - C(i)) \prod_{j=1}^{i-1} C(j) \tag{1}$$

The C/W/L framework assumes that users can only accumulate their gains via the form of *expected total gain* (ETG) or *expected rate of gain* (ERG). Moffat *et al.* [32] extended the C/W/L framework to the C/W/L/A framework by introducing a new component: *Aggregation* (A). Under the C/W/L/A framework [32], the metric score can be computed through the probability a user exits at each rank (L) and the aggregate gain value (A) the user acquires at each rank:

$$M_{\text{CWLA}}(\mathbf{r}) = \sum_{i=1}^{\infty} L(i) \cdot A(i) \tag{2}$$

where $\mathbf{r} =< r_1, r_2, \cdots, r_i >$ is the relevance levels of the documents from position 1 to i, $L(i)$ is the last probability; $A(i)$, *i.e.*, the aggregation function, is how the utility gained by a user from the 1st document to the i-th document is aggregated.

The introduction of aggregation allowed the C/W/L/A framework to characterize metrics like ERR through incorporating an appropriate aggregation function. A new metric can be developed via C/W/L/A framework by simply defining continuation probability and aggregation function. In this study, all metrics are instantiated under the C/W/L/A framework.

In user models of the above metrics, users's decision making is assumed to be affected by multiple factors, including: (1) current position in SERP, (2) the absolute gain they have accumulated from items they have examined so far, or the absolute gain of the current item, (3) the cost (or effort) they have input, and (4) other user characteristics given by external parameters of the metric, such as *persistence* (ϕ) in RBP [35]. Nevertheless, Liu and Han [27] showed that users' search satisfaction and behaviors are associated with *relative gains* and losses to reference points, which is not considered by user models of existing metrics. In this study, we propose a metric framework with a user model considering the influence of reference points on user behavior. As far as we know, we are the first to introduce the reference dependence effect into the user model of evaluation metric.

2.2 Cognitive Biases in Search Interaction

Behind varying user models associated with the above metrics is the premise that users make their decisions rationally according to the absolute gain, cost and maximized expected outcomes in search interaction. However, evidence from cognitive psychology and behavioral economics suggested that one's decisions under uncertainty can systematically deviate from what is expected given rational decision-making models because of *cognitive biases* [50–52]. Cognitive biases arise from one's limited cognitive ability and resources to properly collect and process available information [24]. Apart from psychology experiments, previous studies confirmed that cognitive biases also occur in searching [2, 25].

Recently, some researchers have examined various kinds of cognitive biases that occur in search interaction, such as the ordering effect [1, 10, 17, 18, 45, 56], the bandwagon effect [10, 23], the anchoring effect [46, 49], the recency effect [30], and the reference dependence [27]. Azzopardi [2] summarized the emerging works on cognitive biases in information retrieval and emphasized the importance of integrating human bias features into evaluation.

With the growing knowledge about users' cognitive biases, some recent works began to introduce cognitive biases into the construction and meta-evaluation of evaluation metrics. Zhang *et al.* [59] proposed a metric framework, namely Recency-aware Session-based Metric (RSM), for session-level evaluation. RSM modifies the weights of query-level scores given by metrics like DCG and RBP in order to reflect the recency effect. Chen *et al.* [15] proposed a query-level metric framework modifying existing metrics like DCG, RBP and ERR by incorporating the anchoring effect into the aggregation function.

In this study, we focus on reference dependence effect and develop a *query-level* evaluation metric framework with a user model taking the effect of reference points into consideration. According to Tversky and Kahneman [51, 52], when making decisions under uncertainty, the carriers of value behind one's actions are *gains and losses relative to reference points, rather than final absolute outcomes.* Thus, when people make decisions, their judgements are often influenced by the difference between the associated final outcome and the dynamic reference point at the time of evaluation. Liu and Han [27] found that users' search satisfaction and many aspects of search behaviors and decisions are substantially associated with relative gains, losses and the associated reference points. Taking a step forward, Brown and Liu [7] developed early prediction models based on *simulated initial references*, aiming to address the cold start problem in session behavior predictions. Although the above studies contribute to the integration of cognitive bias concepts with IR problem space, they did not further propose a formal user model that leverages the knowledge learned about reference dependence effect in enhancing evaluation metrics. This leaves a vacuum at the interdisciplinary problem of how to develop an accurate and psychologically more realistic user model for offline metrics by taking user biases into consideration.

2.3 Meta-Evaluation of Evaluation Metrics

With enumerous evaluation metrics developed, the meta-evaluation of evaluation metrics becomes a growing research concern in IR.

User satisfaction is regarded as a near-ideal ground truth metric of retrieval effectiveness and has been applied in meta-evaluating a variety of online behavior-based and offline outcome-based metrics. To measure in what extend metric scores are consistent with users' satisfaction feedbacks, some researchers use correlations with users' satisfaction feedbacks [16, 29, 55, 60], while others use agreements with users' SERP preference [43].

Another widely-used method is *discriminative power* [37]. It is the statistical power of a metric to significantly discriminate system pairs. Discriminative power measures the *stability* of a metric across the topics based on significance testing [38]. It reflects the reliability of a metric. However, it does not tell whether metrics are "measuring what we want to measure" [38] (e.g., how well a metric is correlating with users' satisfaction feedback). Thus it meta-evaluates metrics on a dimension orthogonal to user satisfaction.

Other meta-evaluation methods include swap method [8], judgement cost [11], and so on. In this study, we measure the effectiveness of measures from the perspective of the correlation with users' satisfaction feedbacks and discriminative power as they are widely adopted meta-evaluation methods.

3 REFERENCE-DEPENDENT METRICS

In this section, we introduce the novel setup of our reference-dependent metrics and explain how they relate to existing metrics.

3.1 A Reference-aware User Model

In our reference dependence metric (ReDeM) framework, we define the probability a user continues at rank i as follows:

$$C_{\text{ReDeM}}(i) = \frac{1 + i - r_i}{2 + i - (r_i - r_{\text{ref}})} \tag{3}$$

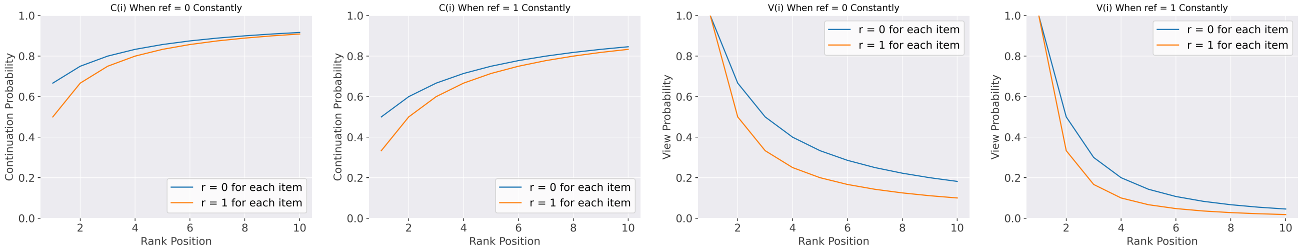

Figure 1: Continuation Probability $C(i)$ and View Probability $V(i)$ of ReDeMs with $r_{\text{ref}} \equiv 0$ and $r_{\text{ref}} \equiv 1$ respectively when relevance scores (r_i) of all items in the SERP are 0 or 1.

where r_i is the relevance level of the i-th item ranging in $[0, 1]$, r_{ref} is the relevance level of the item playing as reference point (See 3.2). This model setting provides the ReDeM with following properties:

- **Adaptive User Behavior**. Similar to ERR and INST, the ReDeM is *adaptive*, which means that the probability of whether a user continues to examine the next item ranked at $i+1$ on the SERP is influenced by the relevance levels of the top i documents. *Ceteris paribus* (assuming all other factors remain the same), as the relevance of the current document r_i increases, $C_{\text{ReDeM}}(i)$ decreases, and *vice versa*. Our model assumes that as r_i increases, users are more likely to feel satisfied and thus end their browsing.
- **Sunk Cost Recovery**. Like DCG, INSQ and INST, $C_{\text{ReDeM}}(i)$ increases as users progress deeper in the ranking, *ceteris paribus*. This is the "sunk cost" property suggested by Moffat *et al.* [31, 34] assuming that when users has already put more effort into the search, they are more likely to continue browsing. The effect of sunk cost is confirmed by previous empirical studies [34, 54] and thus is represented here.
- **Reference Sensitive**. Inspired by cognitive psychology and IR research on reference dependence [20, 27], we argue that users' reference points plays a key role on the user browsing behavior. Specifically, *ceteris paribus*, as r_{ref} increases, $C_{\text{ReDeM}}(i)$ decreases. The behavioral assumption is that: when the gain at the reference point increases, the user is more likely to get less gain than the reference point from subsequent documents ($r_i - r_{\text{ref}} < 0$), which is a "loss" perceived by the user. Due to loss aversion, the user is more likely to end their browsing at current position and settle up their gains. For example, if a user is currently handling the i-th item and the reference point is the average relevance of items from rank 1 to $i-1$, if the average relevance acts as the rate of gain of the user, we would find that the user is more likely to stop as their rate of gain increases. This is consistent with the core arguments of Information Foraging Theory (IFT), which assumes that a point at which a forager should move to the next patch is the point when they achieve the highest gain per unit of cost [4, 6]. With the insights from cognitive psychology, we extended the IFT to cover multiple reference points and developed a flexible multi-stage (e.g., continuation, view, click or stop) framework consistent with existing offline evaluation approaches.

3.2 Selection of Reference Points

To investigate the effect of different reference points on the performance of the framework, we chose the following four simulated reference points based on the findings discussed in psychology and behavioral economics literature [20, 22, 36, 52]:

- **Init**: the relevance level of the first item in the SERP (the anchoring effect).
- **Max**: the best relevance level among the items the user has observed so far. For example, if a user is currently observing the i-th item, the reference point will be the best relevance level among the items from rank 1 to $i-1$.
- **End**: the relevance level of the last element that the user observed. For example, if a user is currently observing the i-th item, the reference point will be the relevance level of the $(i-1)$-th item.
- **Avg**: the average relevance of items the user has observed so far. For example, if a user is currently observing the i-th item, the reference point will be:

$$r_{\text{ref}} = \frac{\sum_{j=1}^{i-1} r_j}{i-1}.$$

- **Peak-End (PE)**: (Ref. Max + Ref. End)/2 (peak-end rule).

To visualize the properties of ReDeMs, Figure 1 shows the Continuation Probability $C(i)$ and View Probability $V(i)$ (the probability that a user will inspect the i-th result supposed the user always starts with the first result in a top-down order) of ReDeMs with $r_{\text{ref}} \equiv 0$ and $r_{\text{ref}} \equiv 1$ respectively when relevance scores (r_i) of all items in the SERP are 0 or 1. From Figure 1 we can see that: (1) $C(i)$ gradually increases as i becomes larger; (2) when r_{ref} is larger, $C(i)$ becomes smaller; (3) when r_i is higher, $C(i)$ becomes smaller; (4) $V(i)$ decreases faster when r_{ref} is larger; (5) $V(i)$ decreases faster when r_i is larger.

4 METHODOLOGY

In this section, we describe the datasets and the experimental design employed in our study. In order to compare performances of metrics in terms of correlation with users' satisfaction feedbacks, we need datasets whose SERPs are labelled with users' satisfaction feedbacks. On the other hand, In order to compare the ability of metrics in terms of discriminating systems with statistical significance, we need a dataset that contains the results returned by different retrieval systems on several topics.

Table 1: Overview of the datasets used in experiment 1 and 2.

	TianGong-Qref [14]	THUIR1 [16]	THU-KDD19 [30]
Data Collection	Naturalistic environment	Controlled lab (Ad hoc retrieval)	Controlled lab (Whole session)
#Sessions	2,356	-	225
#Queries and SERPs	7,479	2,435 (2,391)	1,111
#Results per SERP	≥ 10	10	10
Relevance and Usefulness Judgments	4-level self-rating usefulness scores	4-level external graded judgments	5-level external graded judgments and 4-level self-rating usefulness scores
Query Satisfaction Feedback	5-level ratings	5-level ratings	5-level ratings

Table 2: Overview of the NTCIR-WWW3 [42] dataset used in Experiment 3.

#topics	rel. levels	#rel. per topic	#runs
80	4	159.0	39

To answer **RQ1**, we compare the effectiveness of ReDeMs as well as other ad hoc offline evaluation metrics in terms of the correlation with the user satisfaction feedbacks on TianGong-Qref dataset [14], THUIR1 dataset [16] and THU-KDD19 dataset [30] (refer to 4.2). To answer **RQ2**, we split TianGong-Qref dataset into search scenarios with different *cognitive load* and meta-evaluate the performance of ReDeMs with varying reference points under different search scenarios (refer to 4.3). To answer **RQ3**, we evaluate the discriminative power of ReDeMs and baselines on the NTCIR-WWW3 dataset [42].

4.1 Datasets

The TianGong-Qref dataset [14] contains $2,356$ search sessions, $7,479$ queries and associated SERPs for each query, 4-level user self-rating usefulness scores for all search results on SERPs, and 5-level user satisfaction feedback for each query-SERP pair. For each search session, task information such as task urgency, task difficulty and user expertise (i.e., the extent to which the user is familiar with the search topic), are collected. The user search behavior log of the dataset is collected by a Chrome extension which records users' search-related activities under naturalistic environment.

The THUIR1 dataset [16] is collected in a controlled lab setting. It contains $2,435$ *ad hoc* queries and the SERP for each query, 4-level graded relevance labels given by external assessors for all the 10 results on each SERP, and 5-level user satisfaction for each query.

The THU-KDD19 dataset [30] is collected under laboratory environment in which participants were asked to complete some complex search tasks on commercial search engines. It contains 225 search sessions, $1,111$ queries and associated SERPs, 4-level user self-rating usefulness scores for the items they clicked, 5-level graded relevance labels given by external assessors for the top 5 items and the items clicked by the user, and 5-level user satisfaction feedback for each query [1].

We pre-process the datasets as follows in our experiments: (1) We linearly map the usefulness scores in TianTong-Qref dataset to $r_i \in \{0, 1/3, 2/3, 1\}$. (2) We removed 44 records that can not be parsed in THUIR1 dataset and eventually get $2,391$ records. We linearly map the relevance scores in THUIR1 dataset to $r_i \in \{0, 1/3, 2/3, 1\}$. (3) We exploit usefulness scores in THU-KDD19 dataset as usefulness scores can better reflect users' feedbacks on documents and helps us more in developing metrics effectively predict users' satisfaction feedbacks. We linearly map the usefulness scores in the THU-KDD19 dataset to $r_i \in \{0, 1/3, 2/3, 1\}$.

The NTCIR-WWW3 [42] dataset is from the NTCIR-15 WWW-3 English subtask whose target corpus is clueweb12-B13 (about 50 million web pages) [2]. It includes 80 topics and 39 runs (including 2 baseline runs), with 4-level relevance judgement for documents.

Table 1 introduces the descriptive features of each dataset employed in Experiments 1 and 2, where we evaluate the performance of ReDeMs against that of baseline metrics with fine tuned parameters. Table 2 introduces the dataset used in Experiment 3, where we examine the discriminative power of the proposed new metrics.

4.2 Experiment 1: The Overall Performance of ReDeMs

To figure out the overall effectiveness of ReDeMs, we compare the performance of ReDeMs to widely used metrics in terms of correlation with user satisfaction feedbacks on TianTong-Qref dataset, TianTong-FSD dataset and THU-KDD19 dataset respectively. In this experiment, we use (1) Precision, (2) ERR, (3) (scaled) DCG [35], (4) RBP, and (5) INST as baselines.

The first step is to select an appropriate aggregation function for ReDeMs in order to compute the metric scores of ReDeMs, as Section 3 only gives the continuation function. Here, we arbitrarily assign *expected rate of gain* (ERG) as the aggregation function since all of the above baseline metrics except for ERR can be considered as ERG Metrics [3, 32]. Using the same aggregation function allows a fair comparison to some degree. Future studies can investigate the performance of ReDeMs under various aggregation functions.

The definition of ERG, the "expected utility accumulated per item inspected" [31], is as follows [32]:

$$A_{\text{ERG}}(i) = \frac{1}{V^{+}} \sum_{j=1}^{i} r_i \tag{4}$$

where

$$\frac{1}{V^{+}} = \sum_{i=1}^{\infty} \prod_{j=1}^{i-1} C(j).$$

For each dataset, we randomly split the data into 10 folds in a completely random way regardless the session, using 6 folds as

[1] We observed some inconsistency in the meta-data of the KDD-19 dataset with what the contributors reported. For example, the count of <session> tags are 225, but the contributors reported there are 450 sessions. Also, the relevance levels given by external assessors are ranging from 0 to 4, instead of 0 to 3 reported by the contributors.

[2] https://lemurproject.org/clueweb12/

Table 3: Means of Kendall's τ_b correlation coefficients between metric scores and query-level user satisfaction feedback on three datasets. Bold font indicates the strongest correlation in each dataset. *, ** and * indicates the difference between the mean of a ReDeM and means of all baselines is significant at $p < 0.05$, $p < 0.01$ and $p < 0.001$ level with a Bonferroni correction respectively.**

	TG-Qref	THUIR1	THU-KDD
Precision	0.318	0.203	0.243
ERR	**0.354**	0.255	0.248
DCG	0.330	0.254	0.250
RBP	0.334	0.261	0.249
INST	0.333	0.261	0.250
ReDeM-Init	0.351^{***}	0.261	0.253^{***}
ReDeM-Max	0.346^{***}	0.262	$\mathbf{0.256}^{***}$
ReDeM-End	0.338^{***}	$\mathbf{0.263}^{***}$	0.255^{***}
ReDeM-Avg	0.344^{***}	0.262^{*}	0.255^{***}
ReDeM-PE	0.343^{***}	0.262^{*}	0.255^{***}

training set to tune the parameters of DCG, RBP and INST. The remaining 4 folds is used as test set to compare the performance of all metrics. We repeat this process 50 times for each dataset. For the calibration of DCG, RBP and INST, we conduct a grid search for each metric to find the parameter value that maximizes the correlation between metric scores and user satisfaction feedbacks. That value is then used in test set. For DCG, we search b in $[2, 5]$ with a step of 0.1; for RBP, we search ϕ in $[0.1, 1)$ with a step of 0.05; for INST, we search T in $[1, 20]$ with a step of 1.

Table 3 reports the results of Experiment 1. As it is shown in Table 3, ERR performs the best on TianGong-Qref dataset, outperforming other metrics with a mean correlation of $\tau_b = 0.354$. Despite being inferior to ERR, the proposed metrics still achieve better performance than other baselines with a certain margin. The difference between the mean of each ReDeM and means of all baselines including ERR is statistically significant ($p < 0.001$). When it comes to THUIR1 dataset, ERR is dwarfed by RBP, INST as well as ReDeMs. ReDeM-Init is outperformed by RBP and INST ($p < 0.001$), but other ReDeMs perform well. ReDeM-End, ReDeM-Avg, ReDeM-PE outperform all of the baselines with a mean correlation of $\tau_b = 0.263$ ($p < 0.001$), $\tau_b = 0.262$ ($p < 0.05$) and $\tau_b = 0.262$ ($p < 0.05$) respectively. ReDeM-Max also outperforms all baselines with a mean correlation of $\tau_b = 0.262$, although the edge is not discernible in terms of statistical significance. ReDeMs become the best perfomers on THU-KDD dataset and the edge is statistically significant ($p < 0.001$). ReDeM-Max is the best performer outstripping all baselines with a mean correlation of $\tau_b = 0.256$ ($p < 0.001$), followed by ReDeM-End, ReDeM-Avg and ReDeM-PE who also outperform all baselines with a mean correlation of $\tau_b = 0.255$ ($p < 0.001$), slightly lower than ReDeM-Max. Note that parameters of DCG, RBP and INST have already been tuned before testing, but ReDeMs still achieve similar or even better performance under such condition. This shows the effectiveness of the propose reference-dependent framework.

Regarding reference points, the result shows that ReDeM-Init performs the best on Tiangong-Qref dataset but hobbles on THUIR1 and THU-KDD dataset; ReDeM-End performs the best on THUIR1 dataset but does not perform well on Tiangong-Qref dataset. ReDeM-Max, ReDeM-Avg and ReDeM-PE perform similarly and their performance is relatively stable. This result indicates that the proposed ReDeMs can achieve similar or better performance in correlations with users' search satisfaction. Our **RQ1** is hereby answered.

However, if we ponder the result with the property of the datasets in mind, there is something noteworthy. Compared to baselines, ReDeMs perform the best on THU-KDD dataset, which is collected under laboratory environment where the search tasks are designed to be complex [30], requiring more cognitive loads. This might suggest that ReDeMs perform better under the tasks that require a high-level cognitive load. To examine the hypothesis, we propose RQ2 and conduct the Experiment 2.

4.3 Experiment 2: The Performance of ReDeMs under Different Task States

Liu and Yu [29] reported that the effectiveness of evaluation metrics in terms of correlating with users' satisfaction feedbacks vary significantly across task states. Moffat *et al.* [32] also argued that factors like the nature of the users, the nature of the tasks must be considered when choosing an evaluation metric. Hence, based on the result of Experiment 4.2, we carry out a by-group evaluation on TianGong-Qref dataset to figure out the performance of ReDeMs under different search scenarios.

We first classify queries in TianGong-Qref dataset according to the user-rated task difficulty and task urgency of each session. Queries in a session with "task difficulty" in [0, 1] and "task urgency" in [0, 1] are classified as "Low Cognitive Load"; queries in a session with "task difficulty" in [2, 4] and "task urgency" and in [2, 4] are classified as "High Cognitive Load"; other queries are classified as "Medium Cognitive Load". The number of queries in each classification is reported in Table 4.

We then calculate metric scores of ERR and ReDeMs and compare their performance in terms of correlation with user satisfaction under different task states.

Table 5 reports the performance of ERR and ReDeMs under different task states (i.e., levels of cognitive load involved in the task). As it is shown in Table 3, metrics have higher correlations with users' satisfaction feedbacks when the cognitive load is low or moderate and the correlations are relatively low when the cognitive load is high. compared to ERR, ReDeMs have an edge in correlations with users' satisfaction feedbacks when the cognitive load is high. ReDeM-Init and ReDeM-Max perform the best and the difference among ReDeMs is paltry. In the scenario of medium cognitive load and low cognitive load, ERR performs better. The difference of the performance among ReDeMs is relatively patent compared to the scenario of high cognitive load.

A possible explanation is that, when facing tasks requiring high-level cognitive load where people need to make decisions under counter-intuitive, conflicting or uncertain conditions, cognitive biases are more likely to take place as people need to reduce the amount of information and uncertainty they need to process [21, 50]. For that reason, reference dependence effect might be more manifest

Table 4: The number of queries of each task state. “CL” stands for Cognitive Load.

	Low CL	Medium CL	High CL
#queries	2,193	3,275	2,011

Table 5: Kendall’s τ_b correlation coefficients between metric scores of ReDeMs and query-level user satisfaction feedback on TianGong-Qref dataset under different task states. “CL” stands for Cognitive Load. Bold fonts indicate the highest correlation with users’ satisfaction feedbacks in the task state. All correlations reported are significantly different from zero, with $p \ll 10^{-6}$.

metric	Low CL	Medium CL	High CL
ERR	**0.304**	**0.366**	0.270
ReDeM-Init	0.301	0.353	**0.280**
ReDeM-Max	0.292	0.346	**0.280**
ReDeM-End	0.275	0.339	0.277
ReDeM-Avg	0.287	0.344	0.278
ReDeM-PE	0.287	0.345	0.279

under tasks requiring high-level cognitive load. Nevertheless, this should be examined by empirical studies in the future.

With respect to **RQ2**, our result indicates that ReDeMs can achieve better performance under the tasks that require a high-level cognitive load and the performance of ReDeMs are modest under those tasks only need a low-level or moderate cognitive load. Thus, in real-time evaluation, ussing ReDeMs for tasks that require a high-level cognitive load could be a good choice.

4.4 Experiment 3: The Discriminative Power of ReDeMs

In offline evaluation practice, a metric that tends to significantly discriminate more system pairs is preferred. The ability to significantly discriminate system pairs is called *discriminative power*. To examine the discriminative power of the proposed metrics, we carry out Experiment 3. In our experiment, we first compute the scores of each metric for 39 runs on 80 topics with cutoff $L = 10$ (metrics are computed @10). Here, we set the parameter $b = 2$ for DCG, $\phi = 0.8$ for RBP and $T = 2.25$ for INST. As there are 39 runs, we get $39 * (39 - 1)/2 = 741$ system pairs on 80 topics. For each metric, we have a 80×39 score matrix. We then carry out significance tests for the metrics. In our experiment, we use a distribution-free, randomized version of the paired Tukey HSD test with 1,000 trials [12, 39] via the Discpower tool [3]. The algorithm to obtain Achieved Significance Level (ASL) is given by Carterette [12].

Figure 2 is an ASL curve showing the result. Since ReDeMs perform similarly in the experiment, we only draw the curve of ReDeM-Avg. Metrics whose curves are close to the origin are the ones with high discriminative power, which means that they produce smaller p-values for many run pairs than other metrics do.

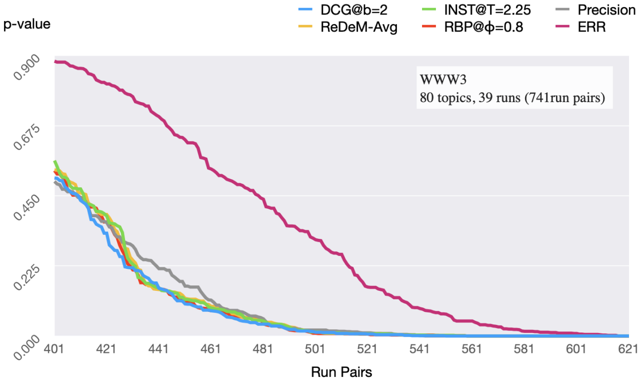

Figure 2: Discriminative power curves of Precision, DCG, RBP, INST and ReDeM-Avg on NTCIR-WWW3 (Randomized Tukey HSD tests for paired data with 1,000 trials).

As it is shown in Figure 2, DCG, RBP and INST perform similarly in terms of discriminative power and the performance of ReDeMs is similar to them. Thus, ReDeMs perform far better than ERR in terms of discriminative power. For example, if a significance level of $\alpha = 0.05$ is required, ReDeMs can discriminate 259 pairs (35% of all) while ERR can only discriminate 186 (25% of all) pairs. The discriminative power of Precision is slightly weaker compared to DCG, RBP, INST and ReDeMs, but is also far higher than ERR. The result shows that ERR has a patently lower discriminative power, which is consistent with what Sakai has observed [40]. Sakai [40] argued that this is caused by the “diminishing return” property of ERR, which assumes that if there is a highly relevant document near the top of the SERP, few users will continue inspecting. Although INST and ReDeMs also have similar property, in their models the impact of a highly relevant document on the continuation probability is relatively smaller.

With respect to **RQ3**, our result indicates that the discriminative power of ReDeMs is far stronger than ERR, slightly stronger than Precision and similar to DCG, RBP and INST.

5 CONCLUSION AND DISCUSSION

This section summarizes the main findings and identifies potential directions for further expanding *bias-aware* IR studies.

In this study, we propose a new evaluation metric framework incorporating the effect of reference points into user browsing models. Our findings suggest that, to build more effective evaluation metrics better reflecting actual user behavior and experience, cognitive factors like the relative gain to a reference point should also be considered besides widely used factors like absolute gain and cost. With growing research efforts on examining algorithmic biases and fairness, our work demonstrates the practical value of integrating the insights from cognitive psychology on *user biases* into user models and evaluation metrics and may inspire future research to keep pushing the boundary on this problem.

The framework proposed can be applied to various scenarios for evaluating search systems and recommender systems, in order to inspire system designs with the awerness of enhance both the

[3] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

effectiveness in online debiasing and the usefulness to boundedly rational users engaging in information-intensive decision-making.

For instance, our model assumes that users will leave earlier due to loss aversion effect if current perceived gain is lower than the reference level. In recommendation scenarios, systems can proactively analyze the item list to be recommended and estimate the potential loss of utility for users because of the cognitive bias triggered by the reference dependence effect. Based on metric scores, systems can adaptively adjust in-situ recommendation strategies accordingly, in order to let users achieve their globally optimal utility.

5.1 Main Findings

With respect to the RQs, we have the following findings:

RQ1: The overall performance of ReDeMs. The proposed ReDeMs with a proper reference point can achieve better correlations with user satisfaction than most existing offline metrics like Precision, DCG, RBP and INST across varying datasets, in some occasions it can achieve better performance than ERR. On TianGong-Qref dataset, ReDeMs outperform all baseline metrics except for ERR; on THUIR 1 dataset, ReDeMs except ReDeM-Init outperform all baseline metrics; and on THU-KDD19 dataset the ReDeMs outperform all baseline metrics. The best performing reference point may vary among different datasets and search study environments.

Although we did not observe huge gains in the experimental result, it is worth noting that the gains are obtained under the condition that DCG, RBP and INST have already been fine-tuned. With these strong baselines, we can confirm the overall effectiveness of our framework in terms of capturing query-level user satisfaction.

RQ2: The performance of ReDeMs under different task states. Experiment 2 shows that compared to baselines, ReDeMs achieve relatively better performance under the tasks that require a high-level cognitive load. This result shows that the actual impacts of reference dependence differ across different search scenarios, and that it is important to investigate user characteristics and task factors in bias-aware search evaluation.

RQ3: The discriminative power of ReDeMs. Experiment 3 shows that the discriminative power of ReDeMs is far stronger than ERR, slightly stronger than Precision and similar to DCG, RBP and INST. As a metric with high discriminative power is preferred in offline evaluation tasks, Experiment 3 shows that ReDeMs with a similar discriminative power to widely used metrics like DCG and RBP could be practically valuable for offline evaluations and may help narrow the gap between simulation-based evaluations and the behavior of boundedly rational users.

In general, the proposed ReDeMs have favorable performance in terms of reflecting users' levels of search satisfaction and discriminating system pairs, which demonstrates the value and potential of our interdisciplinary approach on integrating cognitive bias features into user models and offline evaluation metrics.

5.2 Further Discussion

For the current model, to apply it to a given task, the idea is to adaptively estimate and employ the reference point according to task and user characteristics, which can to some extent be inferred from online behavior features. However, there may be cases where the full details of a search dataset are hard to access beforehand in evaluation practices, or researchers are conducting evaluations on early points of interaction or on new search tasks with very limited interaction data (i.e., cold-start problem in evaluation). Under such cases, choosing theory-informed simulated references [7] might be an acceptable initial solution.

5.3 Limitations and Future Work

Overall, our work illustrates a promising approach to leveraging the insights about user biases from cognitive psychology in better evaluating users' search interactions and demonstrates the importance of modeling cognitive biases for constructing truly person-centered IR systems. There are still many limits remain in this work, which calls for future research efforts on this direction.

The user browsing model we propose is only a preliminary model, which still needs to be improved in order to better reflect users' search behaviors and satisfaction. For example, we did not assign any parameter to the proposed framework. This keeps the simplicity of the framework, but the lack of parameter also limit its adaptability to comprehensively capture diverse user biases, cognitive variations and other contextual restrictions under different task states and user natures. The parameters in DCG, RBP and INST allow them to modify the user model under different task backgrounds and user natures, and thus provide them with a resilience to reflect users' behavior. One direction for the future researches is to incorporate more factors into the model, such as vertical type, browsing order and the interaction effects among different reference points, to make the model more elaborated and realistic in terms of reflecting actual browsing strategies and user perceptions. To empirically support this exploration, more user studies are required in order to explore the impacts of reference points and user biases in general [26, 28].

Experiment 2 shows that ReDeMs have relatively better performance when the cognitive load is high. One of the possible reasons is that cognitive biases are more likely to take place when people facing tasks requiring high-level cognitive load. Thus, reference dependence effect might be more manifest under this case. Future research can continue exploring how the impact of human biases vary across tasks of varying types and the extent to which different biases are correlate with user characteristics and contextual factors.

The proposed framework is designed for and implemented in evaluating *query-level* search interactions only. Liu and Han [27] also showed that the reference dependence effect has influence on the user behavior and the user satisfaction at session level. Based on the measure design and findings from our experiments, future research can further explore *session-level* bias-aware user modelling and search evaluation, and leverage reference dependence effects in understanding and predicting search strategies, transitions of search tactics and cognitive states, as well as overall user experience.

ACKNOWLEDGEMENT

This work was supported in part by NSF IIS-2106152. Any opinions, findings, or conclusions expressed in this material are those of the authors and do not necessarily reflect those of the sponsors. We also thank all reviewers and the meta-reviewer of this paper for their valuable feedbacks and constructive criticism, which are conducive to improve the quality of the study.

REFERENCES

[1] Mustafa Abualsaud and Mark D Smucker. 2019. Exposure and order effects of misinformation on health search decisions. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Rome.*

[2] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/3406522.3446023

[3] Leif Azzopardi, Joel Mackenzie, and Alistair Moffat. 2021. ERR is Not C/W/L: Exploring the Relationship Between Expected Reciprocal Rank and Other Metrics. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (Virtual Event, Canada) *(ICTIR '21)*. Association for Computing Machinery, New York, NY, USA, 231–237. https://doi.org/10.1145/3471158.3472239

[4] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 605–614. https://doi.org/10.1145/3209978.3210027

[5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User Variability and IR System 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, 625–634. https://doi.org/10.1145/2766462.2767728

[6] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* (1989).

[7] Tyler Brown and Jiqun Liu. 2022. A reference dependence approach to enhancing early prediction of session behavior and satisfaction. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–5.

[8] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/345508.345543

[9] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 89–96. https://doi.org/10.1145/1102351.1102363

[10] Keith Burghardt, Tad Hogg, and Kristina Lerman. 2018. Quantifying the Impact of Cognitive Biases in Question-Answering Systems. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). https://ojs.aaai.org/index.php/ICWSM/article/view/15042

[11] Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *SIGIR*.

[12] Benjamin A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Trans. Inf. Syst.* 30, 1, Article 4 (mar 2012), 34 pages. https://doi.org/10.1145/2094072.2094076

[13] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[14] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, M. Zhang, and Shaoping Ma. 2021. Towards a Better Understanding of Query Reformulation Behavior in Web Search. *Proceedings of the Web Conference 2021* (2021).

[15] Nuo Chen, Fan Zhang, and Tetsuya Sakai. 2022. Constructing Better Evaluation Metrics by Incorporating the Anchoring Effect into the User Model. In *2022 International ACM SIGIR Conference on Research and Development in Information Retrieval.* https://doi.org/10.1145/3477495.3531953

[16] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-Evaluation of Online and Offline Web Search Evaluation Metrics *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/3077136.3080804

[17] Tadele T. Damessie, J. Shane Culpepper, Jaewon Kim, and Falk Scholer. 2018. Presentation Ordering Effects On Assessor Agreement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 723–732. https://doi.org/10.1145/3269206.3271750

[18] Michael Eisenberg and Carol Barry. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science* 39, 5 (1988), 293–300.

[19] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[20] Daniel Kahneman. 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review* 93, 5 (December 2003), 1449–1475. https://doi.org/10.1257/000282803322655392

[21] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York.

[22] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives* 5, 1 (March 1991), 193–206. https://doi.org/10.1257/jep.5.1.193

[23] Diane Kelly, Amber Cushing, Maureen Dostert, Xi Niu, and Karl Gyllstrom. 2010. Effects of Popularity and Quality on the Usage of Query Suggestions during Information Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/1753326.1753334

[24] Arie W. Kruglanski and Icek Ajzen. 1983. Bias and error in human judgment. *European Journal of Social Psychology* 13, 1 (1983), 1–44. https://doi.org/10.1002/ejsp.2420130102

[25] Annie Y.S. Lau and Enrico W. Coiera. 2007. Do People Experience Cognitive Biases while Searching for Information? *Journal of the American Medical Informatics Association* 14, 5 (2007), 599–608. https://doi.org/10.1197/jamia.M2411

[26] Jiqun Liu. 2022. Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management* 59, 5 (2022), 103007.

[27] Jiqun Liu and Fangyuan Han. 2020. Investigating Reference Dependence Effects on User Search Interaction and Satisfaction: A Behavioral Economics Perspective. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1141–1150. https://doi.org/10.1145/3397271.3401085

[28] Jiqun Liu and Chirag Shah. 2019. Interactive IR user study design, evaluation, and reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 11, 2 (2019), i–93.

[29] Jiqun Liu and Ran Yu. 2021. *State-Aware Meta-Evaluation of Evaluation Metrics in Interactive Information Retrieval.* Association for Computing Machinery, New York, NY, USA, 3258–3262. https://doi.org/10.1145/3459637.3482190

[30] Mengyang Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Cognitive Effects in Session-level Search User Satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019.* 923–931. https://doi.org/10.1145/3292500.3330981

[31] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Trans. Inf. Syst.* 35, 3, Article 24 (jun 2017), 38 pages. https://doi.org/10.1145/3052768

[32] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A flexible framework for offline effectiveness metrics. In *2022 International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[33] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *ADCS*.

[34] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus Models: What Observation Tells Us about Effectiveness Metrics. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 659–668. https://doi.org/10.1145/2505515.2507665

[35] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (dec 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[36] Donald A Redelmeier and Daniel Kahneman. 1996. Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *pain* 66, 1 (1996), 3–8.

[37] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 525–532. https://doi.org/10.1145/1148170.1148261

[38] Tetsuya Sakai. 2014. *Metrics, Statistics, Tests.* Springer Berlin Heidelberg, Berlin, Heidelberg, 116–163. https://doi.org/10.1007/978-3-642-54798-0_6

[39] Tetsuya Sakai. 2018. Laboratory experiments in information retrieval. *The information retrieval series* 40 (2018).

[40] Tetsuya Sakai. 2021. On the Instability of Diminishing Return IR Measures. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part I.* Springer-Verlag, Berlin, Heidelberg, 572–586. https://doi.org/10.1007/978-3-030-72113-8_38

[41] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 473–482. https://doi.org/10.1145/2484028.2484031

[42] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2021. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task.

[43] Tetsuya Sakai and Zhaohao Zeng. 2021. Retrieval Evaluation Measures That Agree with Users' SERP Preferences: Traditional, Preference-Based, and Diversity Measures. *ACM Trans. Inf. Syst.* 39, 2, Article 14 (dec 2021), 35 pages. https://doi.org/10.1145/3431813

[44] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4 (01 2010), 247–375. https://doi.org/10.1561/1500000009

[45] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. 2013. The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 623–632. https://doi.org/10.1145/2484028.2484090

[46] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and Adjustment in Relevance Estimation. 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, 963–966. https://doi.org/10.1145/2766462.2767841

[47] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2348283.2348300

[48] Richard H. Thaler. 2016. Behavioral Economics: Past, Present, and Future. *American Economic Review* 106, 7 (July 2016), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

[49] Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The Crowd is Made of People: Observations from Large-Scale Crowd Labelling. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 25–35. https://doi.org/10.1145/3498366.3505815

[50] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. http://www.jstor.org/stable/1738360

[51] Amos Tversky and Daniel Kahneman. 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics* 106 (1991), 1039–1061.

[52] Amos Tversky and Daniel Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* 5, 4 (1992), 297–323. http://www.jstor.org/stable/41755005

[53] Ellen M. Voorhees. 2002. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 355–370.

[54] Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical Evidence for Search Effectiveness Models. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (2018).

[55] Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, User Models, and Satisfaction. 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, 654–662. https://doi.org/10.1145/3336191.3371799

[56] Mingda Wu, Shan Jiang, and Yan Zhang. 2012. Serial Position Effects of Clicking Behavior on Result Pages Returned by Search Engines. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 2411–2414. https://doi.org/10.1145/2396761.2398654

[57] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected Browsing Utility for Web Search Evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1561–1564. https://doi.org/10.1145/1871437.1871672

[58] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating Web Search with a Bejeweled Player Model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 425–434. https://doi.org/10.1145/3077136.3080841

[59] Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. *Cascade or Recency: Constructing Better Evaluation Metrics for Session Search*. Association for Computing Machinery, New York, NY, USA, 389–398. https://doi.org/10.1145/3397271.3401163

[60] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 379–388. https://doi.org/10.1145/3397271.3401162